Universidad Autónoma de Madrid
Escuela politécnica superior

Máster Universitario en Bioinformática y Biología Computacional

Trabajo Fin de Máster

Análisis de variación estructural en cáncer mediante secuenciación de lecturas largas

Autor: Alejandro Martín Muñoz
Tutor: Tomás Di Domenico
Ponente: Luis Del Peso Ovalle

Febrero 2021

Análisis de variación estructural en cáncer mediante secuenciación de lecturas largas

Autor: Alejandro Martín Muñoz
Tutor: Tomás Di Domenico
Ponente: Luis Del Peso Ovalle

Unidad de Bioinformática
Centro Nacional de Investigaciones Oncológicas (CNIO)
Febrero 2021

Resumen

El cáncer es una enfermedad compleja, originada como resultado de cambios que pueden darse a distintos niveles del genoma celular. Los avances en las técnicas destinadas a su estudio han abierto las puertas al conocimiento de las mutaciones y los genes implicados en el proceso cancerígeno. Entre ellas, las tecnologías de secuenciación de segunda generación han desempeñado un papel fundamental, al ser capaces de secuenciar el genoma completo de las células presentes en la muestra de un tumor particular. Esto ha permitido la construcción de estándares de referencia tumorales, tanto somáticos como germinales, aunque las mutaciones somáticas constituyen la causa predominante de su desarrollo.

Durante los últimos años, las tecnologías de secuenciación de tercera generación han surgido como solución a algunas de las limitaciones intrínsecas de las anteriores. Entre otras, la capacidad de generar lecturas de kilobases a megabases de longitud ha permitido estudiar de forma adecuada regiones del genoma humano más desconocidas (ricas en GC y repetitivas), cuyas características impedían su correcta resolución con lecturas cortas. Estas regiones suponen una fuente de generación activa de ciertos tipos de variantes, de forma que al menos un 70 % de las variantes estructurales del genoma humano había resultado difícil de caracterizar con lecturas cortas. El desarrollo reciente de estas tecnologías de tercera generación presenta la necesidad de describir sus competencias y obstáculos, de lograr su optimización y de desarrollar aproximaciones bioinformáticas para aprovechar al máximo su potencial.

Con el objetivo de estudiar las capacidades y limitaciones concretas de las tecnologías de secuenciación de ambas generaciones en la detección y el estudio de variantes estructurales en cáncer, este proyecto ha trabajado sobre datos de secuenciación de genoma completo de las líneas celulares COLO829 y COLO829BL (un tipo de melanoma y su contraparte sana). Para ello, se ha desarrollando un flujo de trabajo utilizando cada una de las dos generaciones de tecnología.

Respecto al flujo de trabajo desarrollado basado en lecturas cortas, éste permitió identificar más de un 90 % de las variantes presentes en la referencia somática disponible para COLO829. El flujo de trabajo basado en lecturas largas permitió por su parte identificar más de 30 000 variantes de tipo indel y estructurales. Los resultados obtenidos en este trabajo no sólo han contribuido a demostrar y reforzar el papel clave de las regiones repetitivas en la generación de los tipos de variantes de interés, sino también a la detección específica de tipos complejos. De las variantes identificadas con lecturas largas, destaca una de significancia clínica contrastada en el supresor tumoral *PTEN*. También resultan interesantes otras dos que, aunque no disponen del mismo estatus, permiten plantear hipótesis sobre el cáncer de estudio, al afectar a genes participantes en rutas de potencial relevancia (*TP53TG3B*, *RPH3AL*). De hecho, entre las variantes identificadas y sometidas a una filtración preliminar, 558 de ellas afectan a 243 genes implicados en el desarrollo del cáncer, de forma que las posibles hipótesis a plantear y corroborar podrían ser muchas más.

Finalmente, los cruces entre las variantes identificadas por cada flujo de trabajo permitieron proponer nuevas variantes a validar e incluir en la referencia somática de COLO829. Además, entre ellas se encontraron algunas que afectan a 8 genes implicados en el desarrollo cancerígeno.

Palabras Clave

SVs, indels, cáncer, COLO829, secuenciación, lecturas, longitud, repetitivas, ONT, *Illumina*.

Agradecimientos

Quiero darle las gracias a mi familia, por haberme proporcionado los medios para poder tener la oportunidad de formarme lo mejor posible y por haberme educado en el esfuerzo, la constancia y la ilusión. Fueron mis padres, José Antonio y Ana Belén, los que pusieron la primera piedra y me guiaron para construir todo lo que hoy me define como persona y como profesional. Después, fue mi pareja, Beatriz, quien apareció en mi vida hace 5 años, la cambió por completo, y decidió acompañarme cada día y con cada proyecto.

Gracias a Tomás, por haberme tutorizado durante todos los meses que ha llevado desarrollar este Trabajo de Fin de Máster. Cada ayuda, cada consejo y cada reunión ha servido para para interesarme y profundizar en problemas sin resolver, para hacerme pensar y aprender un poco más cada día, y para lanzar mi carrera en el campo de la Bioinformática.

Por último, gracias a los compañeros de la primera fila del Máster, Diego, Álvaro y Sara. El simple hecho de conocerles ha sido un placer. Además, su ayuda y sus consejos con los distintos problemas tratados a lo largo de todas las asignaturas han sido claves para poder avanzar cada día, para conocer distintas aproximaciones que hoy controlo y soy capaz de ejecutar, y para dar la mejor versión académica de mí mismo. Espero haber sido también lo mejor posible para ellos. Tanto en la vida, como en la ciencia, la colaboración tiene mucho más que hacer y decir que la competencia.

Índice general

Índice de Figuras

Índice de Tablas

1 Introducción

1.1. Motivación del proyecto

El cáncer es una enfermedad compleja debida a alteraciones genéticas o epigenéticas en el ADN, cuyos cambios pueden ser congénitos (germinales) o adquiridos (somáticos) y contribuyen a desarrollar el fenotipo maligno. Aunque estos cambios se producen muchas veces de manera aleatoria y natural (error asociado a la replicación genética), la tasa a la que se producen se ve aumentada por agentes de efecto mutagénico, cuyo origen puede ser interno y externo. Estos agentes van desde una predisposición genética heredada desde los progenitores, a la simple y diaria exposición al ambiente que nos rodea, donde influyen hábitos (sueño, alimentación...), consumo de sustancias nocivas (tabaco) o agentes naturales (luz ultravioleta, virus, envejecimiento...), y cuyo impacto es predominante. A lo largo de toda la vida de un individuo, esto implica un daño continuo sobre el ADN que, la mayoría de las veces, se repara. Sin embargo, aquel daño no reparado da lugar a mutaciones fijas que, dependiendo de qué tipo sean y a qué elemento genómico afecten, pueden tener mayor, menor o ningún efecto (Stratton et al., 2009).

Atendiendo a la teoría de Darwin sobre la selección natural, mutaciones en ciertos genes confieren una ventaja a la hora de proliferar y evadir los mecanismos de control, haciendo que crezcan y sobrevivan de manera más eficaz que otras y, por tanto, que sean seleccionadas positivamente. Esto es esencial en el desarrollo del cáncer, de forma que dichos genes se definen como oncogenes y las mutaciones que les afectan, como mutaciones *driver* o conductoras del cáncer. También hay otras alteraciones que, si bien no son tan esenciales para el desarrollo de la enfermedad, pueden contribuir a la misma, una vez iniciado el proceso cancerígeno, las cuales se definen como pasajeras (Stratton et al., 2009). También hay que destacar: mutaciones que afectan a genes encargados de la protección y reparación del ADN, las cuales favorecen a su vez el desarrollo de un fenotipo mutador; y mutaciones que confieren resistencia a la terapia contra la enfermedad. Todas estas mutaciones no sólo actúan a nivel genético, sino que también pueden darse a nivel epigenético, alterando la estructura de la cromatina y la expresión génica.

La identificación de la primera mutación natural causante del cáncer humano se produjo en 1982 (Reddy et al., 1982; Tabin et al., 1982). Desde entonces, el estudio del cáncer ha avanzado rápidamente, en cuya evolución ha sido fundamental el desarrollo tecnológico y, en particular, la aparición de las técnicas de secuenciación masiva, debido a su rapidez y eficiencia, la gran cantidad de información que proporcionan y la disminución de su coste. En este contexto, la tecnología predominante ha sido la correspondiente a la generación de lecturas de tipo *paired-end*

desarrollada por *Illumina*, que se basa en: extracción y fragmentación de ADN a un determinado tamaño, amplificación del ADN y ligación de adaptadores a los extremos por PCR (pudiendo así hacer un seguimiento y emparejamiento de los fragmentos) y secuenciación por síntesis (hibridación de las moléculas resultantes sobre un soporte (*flowcell*) y determinación de su composición por la identificación individual de cada nucleótido, a medida que se añaden en el proceso de copia de cada molécula, resultando en la formación de una colonia de fragmentos por cada una) (Figura 1.1a). El resultado es la obtención de lecturas de hasta 150 nucleótidos y alta precisión (superior al 99.9 %), emparejadas a otras con orientación opuesta y de las que se encuentran separadas por una cierta distancia (Illumina, Inc., 2021).

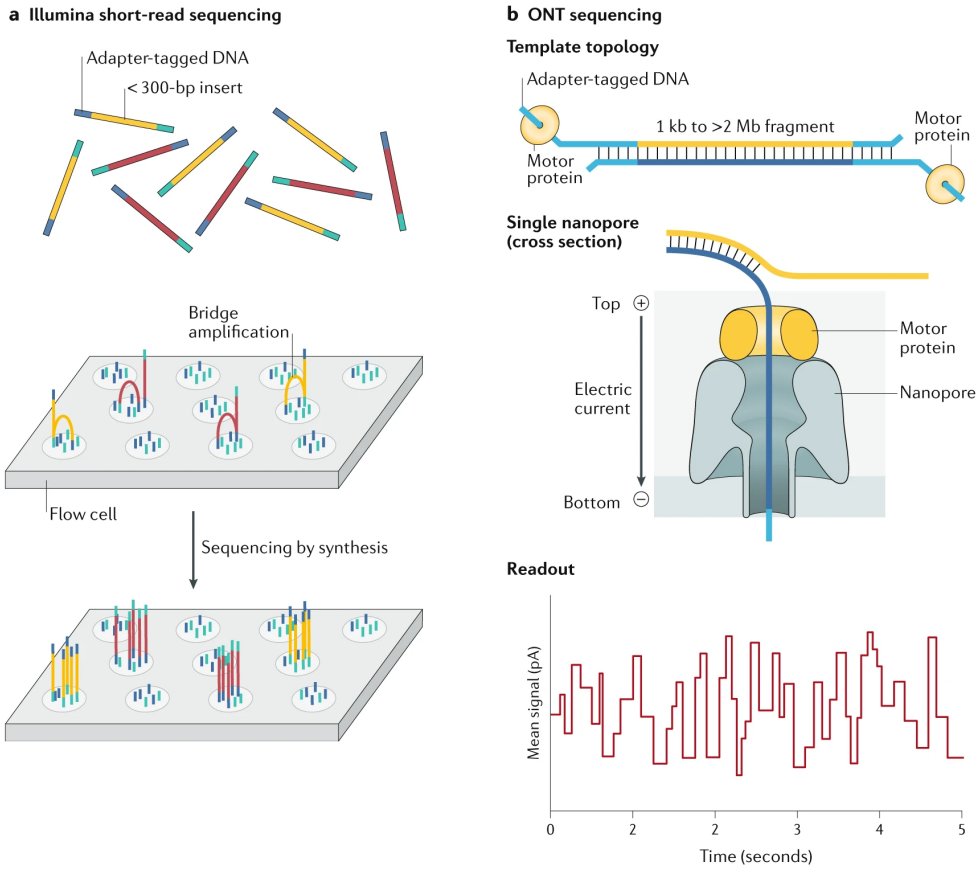

Figura 1.1: a) Visión esquemática de la tecnología de secuenciación de segunda generación y tipo *paired-end*, desarrollada por *Illumina*. **b)** Visión esquemática de la tecnología de secuenciación de tercera generación basada en nanoporos, desarrollada por *Oxford Nanopore Technologies*. Figuras tomadas y modificadas de Logsdon et al., 2020.

A través de herramientas bioinformáticas, estas lecturas pueden utilizarse para identificar biomarcadores relevantes para el estudio de enfermedades como el cáncer, así como para identificar dianas de posibles terapias o predecir la respuesta de los pacientes a ellas. De hecho, esta tecnología ha sido la base de incontables descubrimientos científicos, no solo en el campo de la enfermedad, sino también en otros como el de la evolución, al permitir el estudio de variantes de un nucleótido (*Single Nucleotide Variant*, SNV), variantes del número de copias (*Copy Number Variant*, CNV) y pequeñas inserciones o deleciones (indels) (Craig et al., 2016; Logsdon et al., 2020).

Las variantes estructurales (*structural variant*, SV) comprenden otro tipo de variantes genéticas, definidas como aquellas con un tamaño de al menos 50 nucleótidos, de forma que abarcan desde inserciones (INS), deleciones (DEL), inversiones (INV) y translocaciones de fragmentos de ADN, a diferencias en el número de copias. Como contrapartida a los avances generados por

la secuenciación de segunda generación, su naturaleza supone una limitación para el ensamblaje del genoma (Figura 1.2) y para la detección de más del 70 % de las SVs del mismo, debido a la corta longitud de las lecturas que produce. Como resultado, se desconoce una parte considerable de nuestro genoma (más del 15 %) debido a su naturaleza repetitiva o su alto contenido en los nucleótidos GC (Logsdon et al., 2020). Por otro lado, estas regiones menos estudiadas no solo son de gran importancia a nivel funcional para el genoma sino que, además, son especialmente activas a nivel mutagénico (Sudmant et al., 2015).

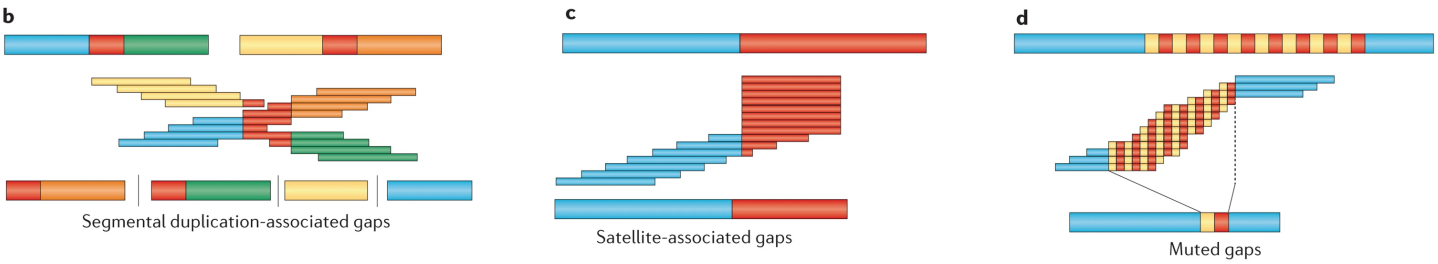

Figura 1.2: Representaciones esquemáticas de ensamblajes genómicos. El genoma a resolver se ubica en la parte superior de cada figura individual como barras gruesas. Las secuencias repetitivas se muestran en rojo. Las superposiciones de lecturas se ubican debajo del genoma como barras finas. Las secuencias ensambladas se muestran en la parte inferior de cada figura individual. Los espacios se muestran como barras verticales e indican secuencias no resueltas. **b)** Dos conjuntos de lecturas comparten una parte repetitiva, lo que impide la ubicación exacta de las lecturas. **c)** Los satélites, regiones repetidas asociadas a áreas centroméricas, acrocéntricas y teloméricas de los genomas, llevan a la creación de lecturas repetitivas que, al no poder distinguirse, se colapsan, no pudiendo definir correctamente la región en la que se encuentran. **d)** Las lecturas generadas en regiones del genoma constituidas por la repetición en tándem de la misma secuencia no consiguen distinguir entre las repeticiones, haciendo que se colapsen. Figura tomada y modificada de Chaisson et al., 2015.

Por ejemplo, hay eventos meióticos en los que se produce una recombinación homóloga desigual entre secuencias repetitivas específicas y flanqueantes de una determinada región, la cual sufre una alteración que viene determinada de forma importante por el número y la orientación de dichas repeticiones (Giglio et al., 2001). También destacan los elementos genómicos de tipo Alu, que constituyen el elemento repetitivo más abundante del genoma humano, llegando a suponer hasta un 10 % de su secuencia. A su vez, estos elementos constituyen la fuente mayoritaria de generación de secuencias repetitivas invertidas. Todos estos elementos provocan inestabilidad genómica, jugando así un papel fundamental como mecanismos de generación de indels y SVs, evidenciados no sólo por eventos de recombinación, sino también por las distintas interacciones que pueden tener lugar entre elementos cercanos o incluso muy lejanos en el genoma (Cook et al., 2020). Estos estudios han podido llevarse a cabo, en parte, gracias a la abundancia de estos elementos, mientras que la contribución de otras secuencias repetitivas o elementos genómicos menos prevalentes se desconoce más, lo que impide disponer de información suficiente para determinar el papel que pudieran estar desempeñando.

A pesar de que se han desarrollado soluciones para sobreponerse a las limitaciones descritas, al estar basadas en lecturas cortas (Lieberman-Aiden et al., 2009; Peters et al., 2012; Zheng et al., 2016), no terminan de solventar todos los problemas, además de que suponen un trabajo y coste mayores.

En este contexto, surgen las tecnologías de secuenciación de tercera generación, basadas en lecturas largas, destacando: *Pacific Biosciences* (PacBio) y *Oxford Nanopore Technologies* (ONT). Estas aproximaciones permiten generar secuencias continuas de kilobases o incluso megabases de longitud, directamente a partir del ADN extraído. Las diferencias entre ellas se encuentran en los enfoques de detección de secuencia y la química, que influyen en las longitudes de lectura, la precisión de las bases identificadas y el rendimiento (Logsdon et al., 2020). En comparación con las tecnologías de secuenciación de segunda generación, el uso de lecturas

más largas permitiría solventar algunas de sus limitaciones intrínsecas. Por ejemplo, en el caso de regiones repetitivas podría generarse un fragmento que no sólo incluyese la parte repetitiva, sino también una parte no repetitiva previa o posterior a la misma. Con ello, se dispondría de un anclaje que permitiría resolver esas regiones complejas, al evitar la ambigüedad en su alineamiento, y estudiar SVs asociadas a ellas.

En 2014 se empezó a comercializar el primer prototipo de un secuenciador basado en nanoporos, *MinION* de ONT. Mientras que otras plataformas utilizan una señal provocada por la incorporación o hibridación de nucleótidos guiados por una hebra de ADN molde, estos secuenciadores detectan directamente la composición de una hebra de ADN. En cuanto a su fundamento, a medida que una corriente eléctrica pasa a través del poro de una proteína (porina), una proteína motora secundaria ayuda a hacer pasar el ADN de interés por dicho poro. Como resultado, el voltaje de la corriente eléctrica se ve afectado, cuyos cambios se registran en el tiempo (Figura 1.1b). Estos cambios son específicos de la secuencia de ADN que atraviesa el poro, pero la precisión de su determinación solo puede llegar al nivel de una combinación de k nucleótidos (k-mero), resultando en muchas señales posibles en lugar de 4. La conversión de las señales de cada k-mero a las bases que los constituyen se lleva a cabo mediante algoritmos de detección de bases, basados actualmente en métodos de aprendizaje automático (Goodwin et al., 2016; Oxford Nanopore Technologies, 2020).

Por supuesto, también se han encontrado problemas asociados a esta tecnología, como la influencia de las modificaciones químicas de los nucleótidos en la señal producida o el rendimiento de los algoritmos de detección de bases, los cuales se han tratado de solucionar. Sin embargo, existen aún problemas pendientes de solución y, seguramente, desafíos actualmente desconocidos (Spealman et al., 2020).

Respecto a la aplicación de las tecnologías de tercera generación, ha aumentado el conocimiento de distintos tipos de variantes genéticas humanas, sobre todo SVs, porque, como ya se ha explicado, sus características diferenciales permiten estudiar regiones e identificar variantes que resultaban inaccesibles para las tecnologías de segunda generación. No obstante, sigue habiendo ciertos tipos de variantes, como las inversiones asociadas a grandes duplicaciones segmentarias, para los que sigue habiendo dificultades (Chaisson et al., 2019).

Estos avances no solo se deben al distinto enfoque ejecutado, sino que van acompañados del desarrollo y la mejora de programas que usan sus datos en los pasos de alineamiento e identificación de variantes. Todos estos acontecimientos no dejan de ser muy recientes de forma que, así como se ha indicado para la tecnología intrínseca a ONT y para los problemas de detección de ciertos tipos de variantes, se sigue necesitando la optimización de los programas ya disponibles y el desarrollo de otros nuevos, para poder seguir avanzando (Logsdon et al., 2020).

1.2. Objetivos y enfoque

Dado el contexto genómico actual, el objetivo de este trabajo es el de definir capacidades y limitaciones concretas de las tecnologías de secuenciación de segunda y tercera generación en la detección y el estudio de SVs en cáncer. Con ello, se pretende abrir el camino al desarrollo de metodologías que permitan explotar el potencial de las lecturas largas para el estudio de este tipo de variantes en enfermedades complejas.

Para alcanzar este objetivo, se ha trabajado sobre datos de secuenciación de genoma completo de las líneas celulares COLO829 (melanoma cutáneo metastático) y COLO829BL (línea linfoblastoide sana del mismo paciente) (Morse and Moore, 1993). En el caso de las lecturas cortas (disponibles para ambas líneas) (Craig et al., 2016), corresponden a una secuenciación de genoma completo a través de la plataforma *Illumina HiSeq2500*, mientras que las largas (dis-

ponibles sólo para COLO829) (Valle-Inclan et al., 2019) corresponden a una realizada por las plataformas *MinION*, *GridION*, y *PromethION* de ONT.

El trabajo sobre cada conjunto de datos implica comprender la complejidad de cada una de las aproximaciones experimentales que los generan. Del mismo modo, este proyecto supone aprender y optimizar el uso de herramientas bioinformáticas para el desarrollo de un flujo de trabajo, con el que usar y tratar los datos de trabajo, destinado a identificar SVs. Este flujo de trabajo incluye: el uso de técnicas de computación de alto rendimiento (*High Performance Computing*, HPC), dada la dimensionalidad de los datos de secuenciación de genoma completo; y el análisis de las variantes genéticas identificadas, a través del cálculo de estadísticas y generación de visualizaciones. Otro objetivo es la consideración del proyecto desarrollado, así como los resultados y análisis derivados del mismo, para extraer y transmitir conclusiones pertinentes. Finalmente, de cara a la posible aplicación del flujo de trabajo creado por hipotéticos usuarios, también se va a proceder al desarrollo y uso de código colaborativo, con el consecuente aprendizaje del *software* de control de versiones asociado al mismo (*git*).

1.3. Metodología y plan de trabajo

Partiendo del estudio de la bibliografía relacionada al tema del trabajo y las técnicas de secuenciación asociadas al mismo, se procedió a elegir los datos válidos con los que plantear y desarrollar el proyecto (Introducción 1.2).

El trabajo sobre cada conjunto de datos se desarrolló haciendo uso de *Snakemake* (Köster and Rahmann, 2012), un sistema de gestión de flujos de trabajo basado en *Python* (Van Rossum and Drake, 2009) que permite: creación de ambientes (*software* y requerimientos) aislados en cada paso ejecutado, pudiendo emplear para ello los recursos del proyecto *Bioconda* (Grüning et al., 2018); ejecución de comandos en *Bash* (Free Software Foundation, Inc., 2019), código escrito en *Python* o programas escritos en *Python* o incluso *R* (R Core Team, 2020). De cara a la aplicación de cualquiera de los flujos de trabajo creados en este proyecto por hipotéticos usuarios, esta aproximación resulta especialmente útil porque, además de que el propio sistema *Snakemake* gestiona por sí solo la instalación de los programas y la definición de los requerimientos que se le indican, se puede ejecutar de forma local o en un sistema HPC, en cuyo caso se encarga también de lanzar los distintos trabajos a la cola de ejecución del mismo.

En Figura 1.3, se puede observar un esquema general de los pasos ejecutados en cada flujo de trabajo creado, uno por cada tecnología de secuenciación y los datos aportados por las mismas, así como los programas empleados en ellos. Ambas aproximaciones se llevaron a cabo usando como genoma de referencia humano tanto la versión GRCh37 como la GRCh38, cuyos resultados han supuesto un consumo de almacenamiento final de aproximadamente 6.3 Terabytes. No obstante, los resultados disponibles en el caso del proyecto en el que se generaron las lecturas cortas corresponden a GRCh37 (Craig et al., 2016). A ello se suma el hecho de que las actualizaciones correspondientes a la nueva versión implican una modificación de las coordenadas genómicas, lo que complica del mismo modo la actualización de los resultados a dicha versión. Como consecuencia, siendo el objetivo del presente trabajo una comparación de las capacidades y limitaciones de cada una de las tecnologías de secuenciación, el análisis final de las variantes obtenidas por cada una, así como las diferentes comparaciones, se ha desarrollado de forma preliminar únicamente para los flujos de trabajo basados en la versión GRCh37 del genoma humano.

Respecto a los análisis posteriores a los flujos de trabajo, las diferentes operaciones, estadísticas, tablas y figuras correspondientes a las variantes identificadas a punto final de los mismos, se obtuvieron con código escrito en *Python* (Materiales y métodos 4.1.5).

El código desarrollado para ejecutar los flujos de trabajo creados puede encontrarse en el repositorio git ubicado en https://gitlab.com/amartin97/cancer_sv. El código desarrollado para llevar a los análisis sobre los datos resultantes de dichos flujos de trabajo puede encontrarse en el repositorio git ubicado en https://gitlab.com/amartin97/cancer_sv_secondary.

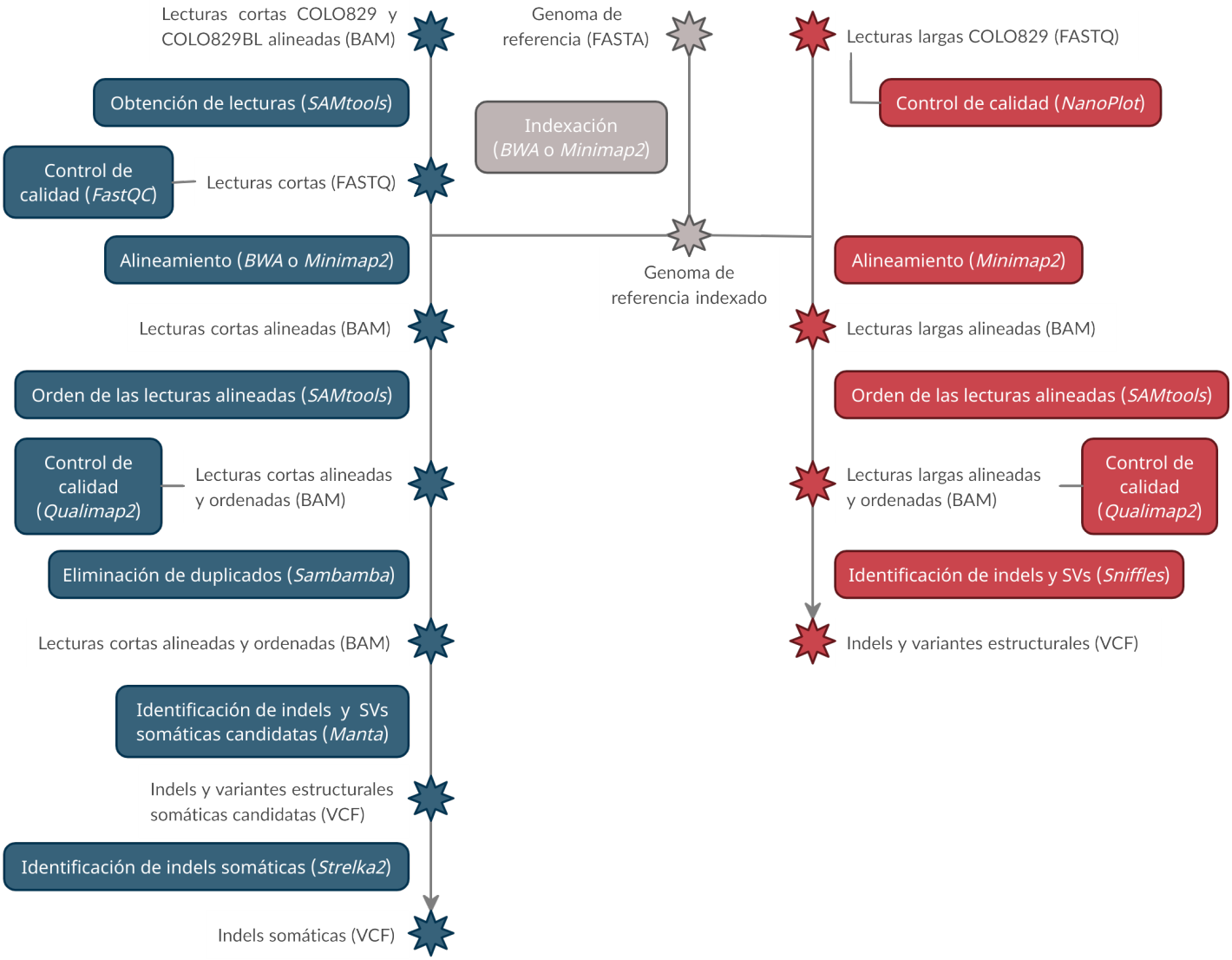

Figura 1.3: Visión esquemática de los flujos de trabajo desarrollados. Los iconos representados con una estrella hacen referencia a archivos, cuyo contenido y formato se indican con el texto adyacente. Los iconos representados con rectángulos hacen referencia a la ejecución de un paso del flujo de trabajo, cuya descripción y programa empleado se muestra en su interior. En la parte izquierda de la figura (y en color azul), se muestra el flujo de trabajo específico de los datos de secuenciación de segunda generación. En la parte central (y en color gris), se muestran pasos y archivos compartidos por ambos flujos de trabajo. En la parte derecha (y en color rojo), se muestra el flujo de trabajo específico de los datos de secuenciación de tercera generación.

2

Resultados

2.1. Flujo de trabajo basado en secuenciación de segunda generación

2.1.1. Análisis general de los resultados

Atendiendo a la referencia de los datos generados por la tecnología de segunda generación (*Illumina*) (Craig et al., 2016) se puede comprobar que se desarrollaron simultáneamente 3 flujos de trabajo independientes, cada uno llevado a cabo por una institución, con sus protocolos y metodologías propios (variabilidad), siendo éstas: *Translational Genomics Research Institute* (TGen), *Canada's Michael Smith Genome Sciences Centre* (GSC), e *Illumina, Inc.* Además, se incorporó un cuarto conjunto de datos, procedente del trabajo que sentó las bases para generar una referencia de variantes somáticas para COLO829 (Pleasance et al., 2010). El resultado del proyecto se sintetiza en un archivo de formato VCF (*Variant Calling Format*) que contiene SNVs e indels, las cuales constituyen la referencia somática de COLO829. A pesar de ello, en este trabajo se decidió intentar la reproducción del análisis llevado a cabo por el proyecto de referencia, partiendo así de los datos de secuenciación originales.

Procedente de cada una de las instituciones citadas, se disponía de sus archivos de alineamiento en formato BAM (*Binary Alignment Map*), conteniendo las lecturas de ambas líneas celulares alineadas frente al genoma de referencia en su versión GRCh37. Sin embargo, dado que trabajar con los datos de secuenciación de las 3 instituciones, y para las dos líneas celulares, supone una cantidad de datos desproporcionada, se decidió emplear únicamente los archivos generados por GSC. La elección se basa en el mayor número de lecturas obtenido por esta institución en comparación con las demás, excluyendo a *Illumina, Inc.* Para poder reproducir el trabajo, se recrearon los archivos de lecturas en formato FASTQ para cada línea celular a partir de los archivos BAM disponibles (Materiales y métodos 4.1.1). Para tener una visión inicial, se llevó a cabo un control de calidad con *FastQC* (Andrews, 2019) de cada archivo disponible, los cuales mostraban una resolución favorable en todos los filtros aplicados y revelaron la existencia de aproximadamente 3 mil millones de lecturas con una longitud de 125 nucleótidos para cada línea celular.

A continuación, se ejecutó el alineamiento de las lecturas *paired-end* con el alineador *BWA* (*Burrows-Wheeler Aligner*) (Li, 2013) (Materiales y métodos 4.1.2). En el caso de las lecturas

largas, se sugiere el uso de otros alineadores optimizados para trabajar con datos de tales características, *Minimap2* (Li, 2018) en este trabajo. Un análisis preliminar del rendimiento de *Minimap2* con lecturas cortas, aplicado a la identificación de variantes de pequeña escala, mostró una rapidez mayor a *BWA-MEM* y una precisión parecida. A diferencia de dicho análisis, este trabajo no está focalizado en el estudio de variantes tan pequeñas como las SNVs, predominantes en la referencia somática, sino en otras de mayor tamaño (indels y SVs). Así, la mayor rapidez y los resultados del análisis preliminar llevaron a preguntarse sobre el rendimiento que podría ofrecer *Minimap2* en el estudio de indels y SVs, de forma que también se procedió a alinear las lecturas *paired-end* con el mismo (Materiales y métodos 4.1.2). En este punto, se ejecutó un control de calidad con *Qualimap2* (comando *bamqc*) (Okonechnikov et al., 2016) sobre los archivos de alineamiento generados por cada programa para cada línea celular, cuyas estadísticas principales pueden observarse en Tabla 2.1. Entre ellas, destaca la ejecución del paso de alineamiento con *Minimap2* en un tiempo inferior a la mitad del empleado por *BWA*. La similitud de los resultados y la gran ventaja en cuanto al tiempo de ejecución llevaron a decidir la inclusión de *Minimap2* como alineador de lecturas cortas, generando entonces dos variantes del flujo de trabajo diferentes sólo en el alineador utilizado.

	BWA + SAMTools		***Minimap2 + SAMTools***	
Tiempo de ejecución (horas)	31		12.5	
Muestra	COLO829	COLO829BL	COLO829	COLO829BL
Tamaño de la referencia	3 234 834 689			
# de lecturas alineadas	2 749 872 415 (97.54 %)	2 883 426 163 (97.16 %)	2 740 644 926 (96.24 %)	2 869 609 186 (95.76 %)
# de lecturas no alineadas	69 476 745 (2.46 %)	84 184 217 (2.84 %)	107 046 374 (3.76 %)	127 111 930 (4.24 %)
Longitud de lecturas mín/máx/media	30 / 125 / 125	30 / 125 / 125	25 / 125 / 124.4	25 / 125 / 124.4
Profundidad de secuenciación media	103X	108X	103X	108X

Tabla 2.1: Tiempo de ejecución del alineamiento, ordenamiento y compresión del archivo de formato SAM resultante en el flujo de trabajo de lecturas cortas, obtenido mediante la ejecución del comando *seff* en el clúster. A continuación, resumen generado por el control de calidad ejecutado con *Qualimap2*. Cada columna muestra los datos correspondientes para la aproximación seguida y, en caso de que sea pertinente, de forma específica para la muestra considerada.

El siguiente paso consiste en el marcaje y exclusión de las lecturas duplicadas, lecturas que son copias idénticas de otras. Atendiendo al proceso de secuenciación llevado a cabo por *Illumina* (Introducción 1.1), la amplificación por PCR durante la preparación de la librería implica la duplicación del material en sí por definición, que además puede ser favorecida para unos fragmentos sobre otros, por ejemplo, aquellos de menor longitud, introduciendo un sesgo. Durante la secuenciación propiamente dicha, la situación ideal sería que una molécula de cada región analizada del genoma generara una colonia en la superficie del *flowcell*, lo que se trata de obtener en función de la concentración de la muestra de ADN empleada, por una cuestión probabilística. Sin embargo, varias copias de la misma molécula original podrían hibridarse, generando dos colonias diferentes y dando lugar a lecturas duplicadas. Todos estos problemas técnicos asociados a la secuenciación por *Illumina* pueden dar lugar a duplicados y, aunque sus plataformas suelen tratar de manejarlos, se sigue requiriendo un marcaje de los mismos con

posterioridad a la secuenciación. En el caso de no tratar la presencia de duplicados, los resultados contendrían sesgos y errores que se derivarían a las conclusiones. En este trabajo, este paso se ha ejecutado mediante el programa *Sambamba* (Tarasov et al., 2015) (Materiales y métodos 4.1.3).

Con el conjunto de alineamientos resultante, se llevó a cabo la identificación de indels somáticas con *Strelka2* (Kim et al., 2018) (Materiales y métodos 4.1.4). El hecho de no poder identificar SVs se debe a que se definen como aquellas con una longitud superior a 50 nucleótidos de forma que, al presentar *Strelka2* un límite de longitud máximo para las variantes identificadas de 49 nucleótidos, solo puede identificar SNVs e indels.

Finalmente, se desarrolló el análisis de las variantes detectadas a punto final del flujo de trabajo (Materiales y métodos 4.1.5). La referencia somática de COLO829 está formada por 35 989 variantes, de las cuales 424 son indels. La reproducción del proyecto de generación de esa referencia (Craig et al., 2016), desarrollada en este trabajo, lleva a la identificación de 128 978 indels somáticas. Con el propósito de validar la aproximación desarrollada, se procedió a realizar una búsqueda estricta de las 424 variantes de la referencia entre 128 919 de las identificadas: una variante es encontrada sólo si hay una en el archivo de formato VCF creado con la que coincida en el cromosoma, la posición en él, los nucleótidos presentes en la referencia y aquellos presentes en la variante.

La causa detrás del uso de un subconjunto con 59 variantes menos de las identificadas en el flujo de trabajo requiere comprender qué componentes constituyen la secuencia empleada como referencia: secuencias de las que se conoce el cromosoma al que pertenecen y su posición en él (*primary assembly*); secuencias de las que se conoce el cromosoma al que pertenecen y su posición en él, pero que constituyen una representación alternativa de un locus del *primary assembly* (*alt scaffolds*); secuencias de las que se conoce el cromosoma al que pertenecen, pero sin poder determinar la posición u orientación que presentan en él (*unlocalized scaffolds*); secuencias de las que se desconoce el cromosoma al que pertenecen (*unplaced scaffolds*); y secuencias que corrigen aquella presente en el *primary assembly*, añaden secuencia a la misma o reducen huecos presentes en la misma (*patches*). Como se puede comprobar, aquellas variantes identificadas en *unlocalized/unplaced scaffolds* deben ser filtradas, puesto que el valor de posición que puedan mostrar no corresponde al real, pudiendo llevar a confusión.

El resultado fue la identificación de 404 de las 424 (95.28 %) variantes presentes en los datos de referencia. No obstante, el análisis de las 20 no identificadas reveló que 2 de ellas sí estaban realmente en el archivo de formato VCF creado, las cuales no pudieron ser encontradas por no coincidir exactamente en los nucleótidos reconocidos como presentes en la referencia (Tabla A.1). Es decir, la reproducción desarrollada consiguió identificar 406 de las 424 (95.75 %) variantes descritas en la referencia somática de COLO829.

En contraste, el número de variantes detectadas pero no presentes en los datos de referencia es de 128 515, un 99.7 % del total tras la filtración. Esto llevó a plantear la aplicación de un filtro adicional con el que disminuir la cantidad de variantes identificadas a una más manejable. Atendiendo al campo *FILTER* del archivo de formato VCF obtenido, se puede comprobar la existencia de los siguientes valores posibles: *HighDepth*, *LowDepth*, *LowEVS*, *HighDepth;LowEVS*, *LowDepth;LowEVS*, *HighDepth;LowEVS;LowDepth*, *PASS*. EVS (*Empirical Variant Score*) corresponde a la probabilidad de que una variante sea un falso positivo, expresada en la escala *phred* y calculada a partir de la aplicación de un *random forest* por *Strelka2*. Así, en el caso de que el EVS de una variante esté por debajo de un umbral establecido, ésta se califica como *LowEVS*. Respecto a los valores relativos a *Depth*, indican si la profundidad de lectura en el locus correspondiente es menor (*Low*) o mayor (*High*) de un cierto umbral. Por último, *PASS* es el valor dado a las variantes que han superado todos los filtros aplicados por *Strelka2*. Por tanto, se procedió a filtrar el archivo de formato VCF para mantener aquellas variantes cuyo campo *FILTER* fuera *PASS* o contuviera el texto *HighDepth*, pero que no contuviera el texto

LowDepth. Después, al igual que se hizo anteriormente, se mantuvieron aquellas que no pertenecieran a *unlocalized/unplaced scaffolds*. Como consecuencia, se obtuvo un subconjunto de 1 411 variantes. La búsqueda estricta de las 424 variantes de la referencia en el subconjunto resultante derivó en 387 variantes encontradas (91.27 %), a las que habría que añadir 1 de las adicionalmente identificadas en la búsqueda anterior, logrando recuperar un total de 388 de las 424 (91.51 %).

En conclusión, la reproducción de la identificación de indels somáticas desarrollada en este trabajo, usando un solo conjunto de lecturas (de cuatro totales) obtenidas mediante secuenciación de tipo *paired-end* con *Illumina* y dos pasos de filtración, ha conseguido recuperar 388 variantes de las 424 de la referencia (91.51 %). Por tanto, este resultado implica identificar más de un 90 % de las variantes presentes en la referencia, usando una cantidad de recursos dramáticamente inferior a la empleada para la creación de la referencia. Esto constituye una prometedora validación del flujo de trabajo creado.

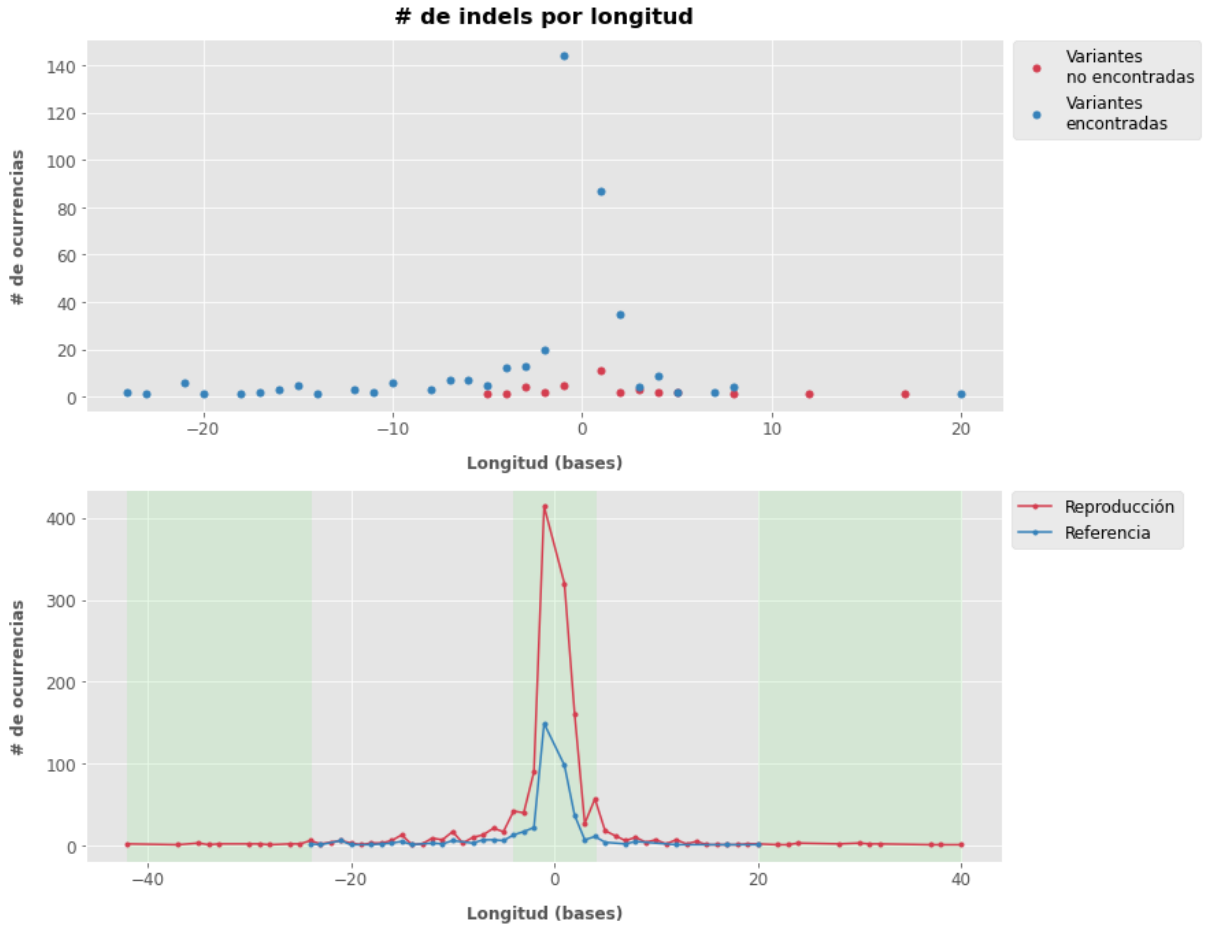

Figura 2.1: Número (#) de variantes por cada valor de longitud posible. Los valores negativos de longitud corresponden a deleciones. En el panel superior, los conjuntos representados corresponden a las variantes de la referencia somática de COLO829 encontradas (azul) y no encontradas (rojo) en el subconjunto de 1 411 variantes resultante del flujo de trabajo de lecturas cortas tras las filtraciones descritas. En el panel inferior, los conjuntos representados corresponden a todas las variantes de la referencia somática (azul), y todas las del subconjunto de 1 414 variantes obtenido sin incluir la filtración de acuerdo a la secuencia en la que se ubican (rojo). Las regiones sombreadas en verde corresponden a intervalos de longitud de interés, resaltados en el texto.

A continuación, se procedió a comparar las longitudes de las indels de la referencia encontradas frente a las de las no encontradas (Figura 2.1, panel superior), con la intención de comprobar la existencia de algún patrón diferencial. La figura resultante parece indicar una mayor dificultad

para identificar inserciones mayores de 10 nucleótidos. Además, llaman la atención las longitudes máximas de dichas indels. Si bien *Strelka2* permite identificar indels en el intervalo de longitud ± 50, el intervalo mostrado por las variantes presentes en la referencia es aproximadamente (-25, 20). Por consiguiente, se procedió a comparar las longitudes de las variantes identificadas por el proceso de reproducción, incluyendo sólo la filtración de acuerdo al campo *FILTER* (1 414 variantes), frente a las de las presentes en la referencia (Figura 2.1, panel inferior). Como se puede comprobar, hay un considerable número de variantes que quedan fuera del intervalo (-25, 20) (regiones sombreadas en verde en los extremos, Figura 2.1, panel inferior), no llegando así a ser incluidas en la referencia. Del mismo modo, hay un gran número de variantes identificadas en el intervalo (-4, 4) (región sombreada en verde en el centro, Figura 2.1, panel inferior) que también son descartadas.

El criterio de inclusión de una variante en la referencia somática consistía en ser identificada por las 4 aproximaciones empleadas o, al menos, por 3 de ellas y que en la 4 hubiera una profundidad de secuenciación insuficiente ($<$ 20 lecturas) (Craig et al., 2016). De acuerdo a este criterio, Figura 2.1 (panel inferior) podría sugerir: una mayor variabilidad y dificultad en la identificación de variantes fuera del intervalo (-25, 20), al identificarse una menor cantidad y no coincidir entre aproximaciones; y una alta tasa de falsos positivos a menor tamaño de la variante, al haber una mayor facilidad de que errores de secuenciación o técnicos puedan llevar a tal identificación. La mayor facilidad de detectar deleciones y menor para inserciones es consistente con los resultados de trabajos previos (Nattestad et al., 2018; Gong et al., 2020).

Como ya se explicó, el proceso descrito se llevó a cabo de dos formas, únicamente diferenciadas por el alineador empleado: *BWA* o *Minimap2*. Así, la reproducción desarrollada con la aplicación de *Minimap2* dio lugar a la identificación de 128 406 indels somáticas, de las cuales se mantienen 128 373 tras filtrar por las secuencias en la que se encuentran. Al realizar la búsqueda estricta de las 424 variantes de la referencia entre ellas, se observó la recuperación de 403 de las 424 (95.05 %), solo una menos que en caso de usar *BWA*. Nuevamente, el análisis de las 21 variantes no identificadas reveló que 2 de ellas sí estaban realmente en el archivo de formato VCF creado, las cuales resultaron ser las 2 mismas descritas anteriormente. Es decir, la reproducción desarrollada, alineando con *Minimap2*, consiguió identificar 405 de las 424 (95.52 %) variantes descritas en la referencia somática de COLO829, apareciendo todas ellas entre las 406 identificadas alineando con *BWA*. En caso de llevar a cabo los dos pasos de filtración descritos, la cantidad de variantes obtenidas se reduce a sólo 1 288 variantes y, al buscar de forma estricta las 424 de la referencia entre ellas, el resultado fueron 386 variantes encontradas (91.04 %), a las que nuevamente habría que añadir 1 de las adicionalmente identificadas en el caso anterior, logrando recuperar un total de 387 de las 424 (91.27 %).

En conclusión, la reproducción de la identificación de indels somáticas desarrollada en este trabajo, usando un solo conjunto de lecturas (de cuatro totales) obtenidas mediante secuenciación de tipo *paired-end* con *Illumina*, usando *Minimap2* como alineador y dos pasos de filtración, ha conseguido recuperar 387 variantes de las 424 de la referencia (91.27 %). Por tanto, este resultado implica identificar más de un 90 % de las variantes presentes en la referencia. Esto supone una prueba más a favor de la validación del flujo de trabajo creado, además de un argumento a favor del reemplazamiento de *BWA* por *Minimap2* en el paso de alineamiento, al suponer la identificación de prácticamente las mismas variantes, pero en un tiempo de ejecución del alineamiento de menos de la mitad (Tabla 2.1).

2.1.2. Análisis de los resultados obtenidos con *Manta*

A diferencia de *Strelka2*, *Manta* (Chen et al., 2016) no solo identifica indels, sino también SVs. Con el objetivo de tener una visión más apropiada de ello, se procedió a representar el número de variantes detectadas por *Manta* en función de su longitud (Figura 2.2, panel superior). En

el caso estudiado, *Manta* identifica 143 144 variantes somáticas, cuyas longitudes abarcan el intervalo (-1 000, 1 000). Centrando la atención en el intervalo de longitudes que es capaz de comprender *Strelka2*, al igual que en su caso, agrupa un considerable número de las variantes pero, en lugar de observarse un pico en el intervalo de longitudes (-4, 4), se resalta uno en el intervalo (-8, -24) y otro en el intervalo (8, 24) (Figura 2.2, panel inferior).

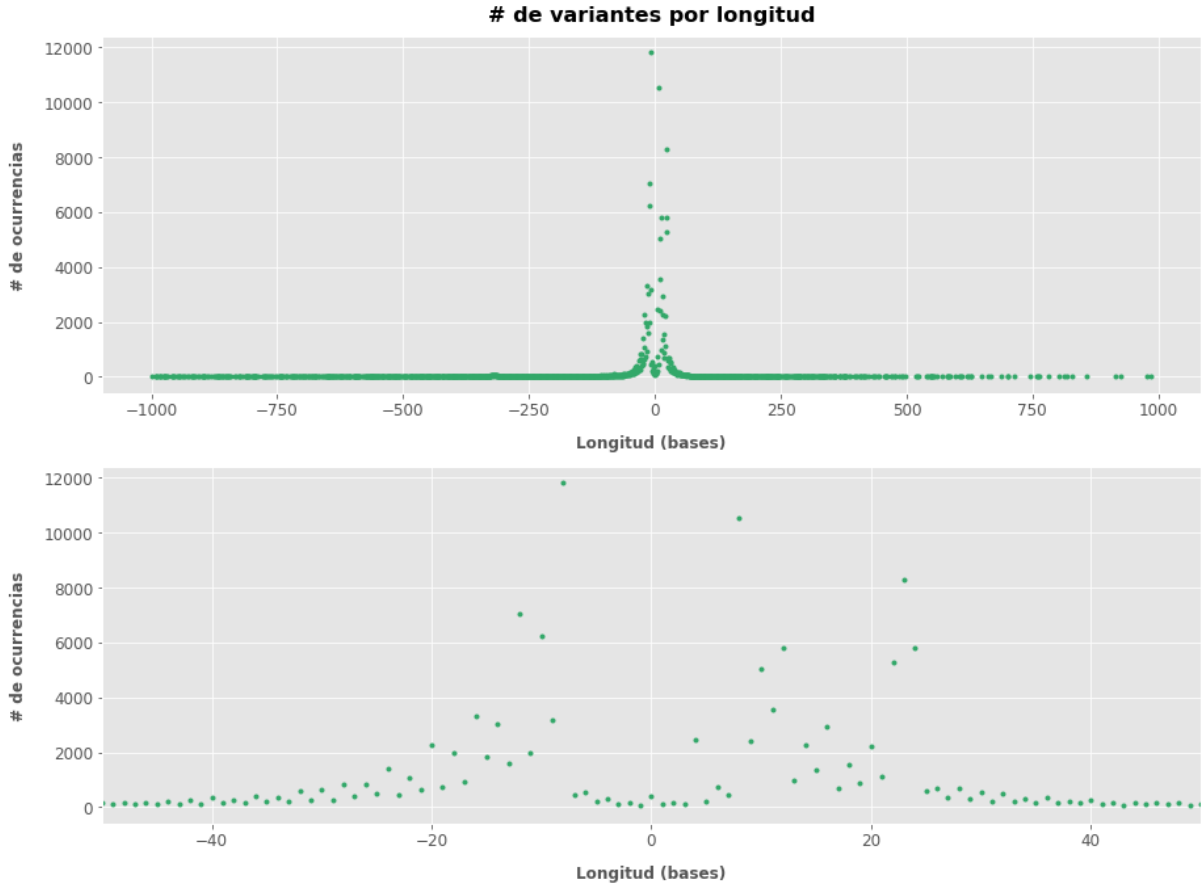

Figura 2.2: Número (#) de variantes por cada valor de longitud posible. Los valores negativos de longitud corresponden a deleciones. En ambos paneles, el conjunto representado corresponde a todas las variantes resultantes del flujo de trabajo de lecturas cortas, hasta la ejecución de *Manta*.

Otra diferencia destacable es el hecho de que *Manta* no solo identifica inserciones y deleciones, sino que también reporta duplicaciones, inversiones, y variantes de tipo BND. En el manual de especificación del formato VCF (Danecek et al., 2011), las BNDs se definen como SVs cuyos puntos de ruptura dificultan su clasificación en las formas canónicas de SVs. En un trabajo previo a éste (Abel et al., 2020), su análisis en profundidad permitió determinar que: algunas eran deleciones de menos de 100 nucleótidos con una profundidad de secuenciación insuficiente; otras, inserciones de retrogenes, causadas por la acción de retroelementos sobre los ARN mensajeros (ARNm); otras, reordenamientos genómicos complejos con múltiples puntos de ruptura; y otras, alteraciones intra o intercromosómicas cercanas o distantes, apoyadas por una baja evidencia. En otro trabajo (Gong et al., 2020), se determinó que dos pares de BNDs pueden representar la inserción de la copia de un fragmento de ADN de una región del genoma en otra o una translocación. Además, resaltan que, a menudo, variantes correspondientes a tipos representados por BNDs se clasifican incorrectamente como deleciones o duplicaciones, disminuyendo la identificación de tipos representados por BNDs y enriqueciendo otros con falsos positivos.

Ambos trabajos citados sugieren la participación de las regiones complejas del genoma (repetitivas) en la generación de este tipo de variantes, complicando su identificación con lecturas cortas y llevando a que hayan sido ignoradas normalmente. No obstante, tampoco hay que olvi-

dar el enriquecimiento con falsos positivos provocado por estas regiones. Como consecuencia, la dificultad y requerimiento de este tipo de variantes ha impedido disponer de tiempo suficiente para poder ser incluidas en el análisis desarrollado por este trabajo.

Para conocer de forma más explícita la contribución de *Manta* en la identificación de variantes somáticas, se procedió a realizar la búsqueda estricta de las 424 variantes de la referencia entre las 123 798 identificadas y filtradas de acuerdo a la secuencia en la que se encuentran y al tipo de variante, al excluirse las BNDs, recuperando 41 de ellas (9.67 %). En este caso, al ser el número de variantes no encontradas mucho mayor, se diseñó otra aproximación para tratar de comprobar la posible existencia de más variantes que realmente sí coincidieran con las de la referencia o que sugirieran una posible contribución en su identificación. Dicha estrategia consistió en buscar aquellas variantes identificadas por *Manta* que coincidieran con las de la referencia en el cromosoma y presentaran una superposición en cuanto a sus coordenadas. Adicionalmente, para aumentar la sensitividad, se consideró como superpuestas a variantes no perfectamente coincidentes, pero con coordenadas dentro de un determinado intervalo de distancias alrededor de la región estudiada. En el primer caso, se consiguió recuperar una variante más. En cuanto a los intervalos, se observó que, usando un intervalo de la posición de la variante en la referencia $\pm$ 15, se recuperaban 6 variantes más con respecto a las 41 identificadas de forma estricta. Finalmente, en un intervalo de la posición de la variante en la referencia $\pm$ 50, solo se añadía una más, a parte de las 6 ya indicadas. Sin embargo, esta última se descartó debido a que las variantes coincidentes no consistían en el mismo tipo de variante (deleción para la identificada por el flujo de trabajo, inserción para la presente en la referencia) y estaban separadas por una distancia considerable, de forma que la existencia de una relación parecía improbable.

Respecto al análisis comparativo de las 6 variantes identificadas con respecto a las coincidentes en la referencia, se encuentran casos como los ya indicados (variantes que caen en la misma posición, pero presentan cierta variación entre los nucleótidos expuestos como de referencia y alternativos, Tabla A.1) y otros que parecen sugerir la existencia de algún tipo de variación, la cual no consigue identificarse de forma inequívoca. A ello hay que añadir que, de los 6 casos a los que se hace referencia, 5 de ellos corresponden a regiones afectadas por secuencias repetitivas, uno de los cuales se muestra en Figura 2.3.

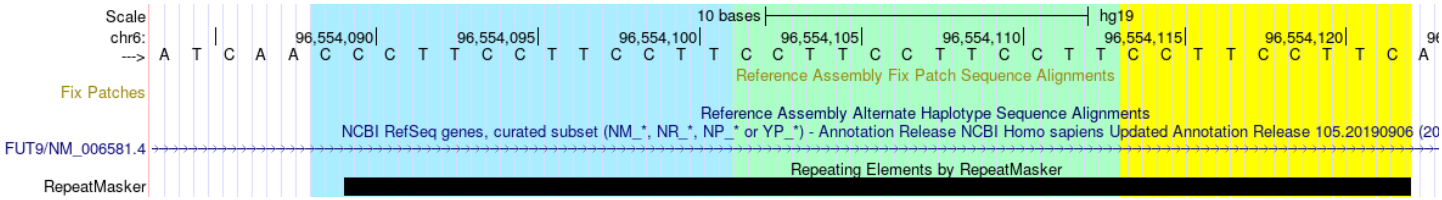

Figura 2.3: Captura del navegador genómico de la Universidad de California en Santa Cruz (UCSC), centrado en una región del cromosoma 6 afectada por una de las variantes estudiadas. La región sombreada en azul y verde corresponde a la afectada por la variante de interés, presente en el conjunto resultante del flujo de trabajo de lecturas cortas hasta la ejecución de *Manta*, incluyendo las filtraciones descritas. La región sombreada en verde y amarillo corresponde a la afectada por la variante de interés, presente en la referencia somática. Así, la región sombreada en verde corresponde a aquella solapante entre las dos variantes estudiadas. El rectángulo negro indica la región correspondiente a un elemento genómico repetitivo.

Estos resultados vuelven a evidenciar la limitación de usar lecturas cortas en tales regiones pues, si bien son 5 variantes incluidas en la referencia, la variante en ella y la identificada por *Manta* no son la misma, lo que sugiere que la variante real podría ser distinta a la reportada. En este punto, tampoco hay que olvidar todas las demás variantes identificadas por *Manta* que no llegan a incluirse en esa referencia, quizás por no tener una evidencia suficiente que las respalde al ubicarse en regiones de este tipo.

2.1.3. Análisis de la influencia de las regiones repetitivas del genoma

Dada la contribución de las regiones repetitivas del genoma como mecanismo generador de indels y SVs (Introducción 1.1), se procedió a buscar qué variantes de las identificadas en la reproducción del proyecto de creación de la referencia somática para COLO829 caen en dichas regiones. Para ello, se hizo uso del archivo de anotaciones de las regiones repetitivas del genoma humano en su versión GRCh37, creado por *RepeatMasker* (Smith et al., 2015). El proceso de búsqueda consistió en seleccionar aquellas variantes que tuvieran parte de su secuencia superpuesta con regiones anotadas como repetitivas (Figura 2.4) (Materiales y métodos 4.1.5).

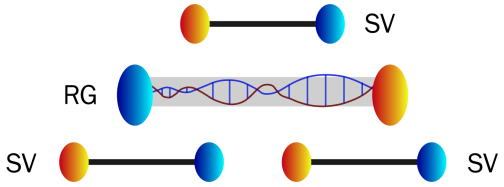

Figura 2.4: Visión esquemática del proceso de búsqueda de intersecciones entre variantes (SV) y genes o regiones repetitivas del genoma (RG). Los extremos de las variantes y región indicados con el mismo color resaltan las posiciones comparadas en cada búsqueda.

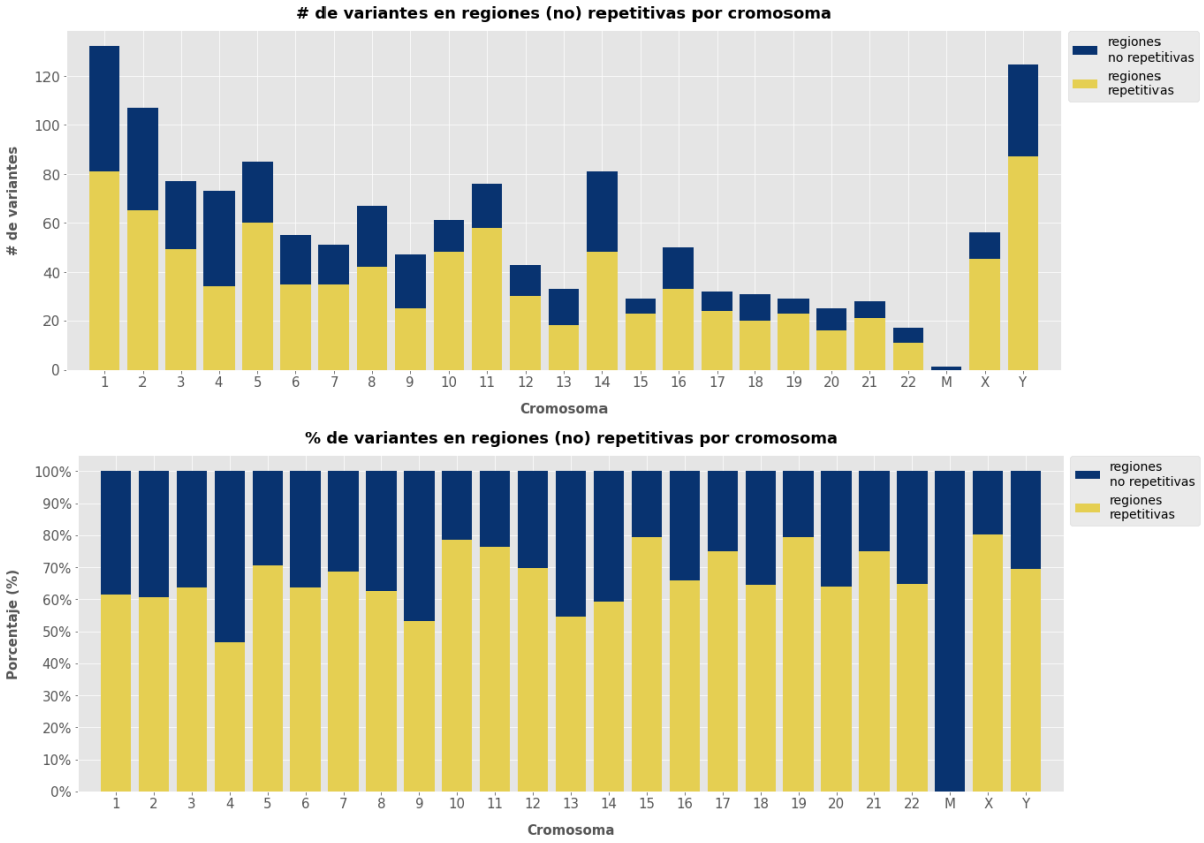

Figura 2.5: Gráficos de barras apiladas representando, para cada cromosoma, el número (#, panel superior) o porcentaje (%, panel inferior) de variantes identificadas por lecturas cortas en regiones no repetitivas en azul y el de aquellas en repetitivas en amarillo.

En este proceso de búsqueda, es necesario volver a tener en cuenta los componentes constituyentes de la secuencia empleada como referencia, de forma que se vuelven a excluir aquellos de los que no se puede determinar el cromosoma al que pertenecen o su ubicación/orientación en él (*unlocalized/unplaced scaffolds*). El conjunto de variantes identificadas por *Strelka2* se somete así al paso de filtración basado en el valor del campo *FILTER* del archivo de formato VCF y a otro

en el que se excluyen aquellas variantes que caen en las secuencias no consideradas, resultando en 1 411 indels.

En el caso del archivo obtenido de *RepeatMasker*, las regiones anotadas como repetitivas constan, entre otras anotaciones, del cromosoma en el que se encuentran (chr1, chrM, chrX, etc.) y de las posiciones inicial y final en él. Como ya se ha explicado, se excluyen aquellas regiones ubicadas en *unlocalized/unplaced scaffolds*, al no ser fiables ni reales las posiciones que puedan presentar. Sin embargo, a diferencia del genoma de referencia, este archivo incluye regiones repetitivas presentes en otros haplotipos, las cuales no se excluyen por considerar que ejercen un papel equivalente al de las secuencias de tipo *alt scaffolds*.

Los resultados observados en Figura 2.5 ponen de manifiesto la gran contribución de las regiones repetitivas en la generación de indels identificadas en COLO829. El porcentaje medio de variantes en regiones repetitivas es del 64.3 % (mediana, 64.71 %), siendo el mínimo de variantes en regiones repetitivas del 46.58 % (excluyendo el genoma mitocondrial, con un 0 %) y el máximo un 80.36 %. Los dos paneles de Figura A.1 también muestran respectivamente la longitud y el porcentaje de cada cromosoma correspondiente a regiones repetitivas y no repetitivas (los cálculos realizados no han incluido las secuencias correspondientes a: *alt scaffolds* ni a diferentes haplotipos, ya que incrementarían el tamaño de cada cromosoma y llevarían a una representación errónea; *unlocalized/unplaced scaffolds*, por las razones descritas).

2.2. Flujo de trabajo basado en secuenciación de tercera generación

2.2.1. Análisis general de los resultados

Respecto al flujo de trabajo relativo a las lecturas largas, solo se dispone de los resultados de la secuenciación de genoma completo con tecnologías de tercera generación (ONT) para COLO829 (Valle-Inclan et al., 2019). La ausencia de los mismos resultados para la contraparte sana (COLO829BL) impide así identificar las variantes somáticas por separado. Esto no supone un problema significativo de cara a la comparación con las variantes presentes en la referencia o las resultantes del flujo de trabajo desarrollado con lecturas cortas, pero sí en cuanto a la detección de nuevas variantes, al no poder confirmar su origen. Por tanto, se debería disponer de los resultados de secuenciación con ONT de la muestra normal para confirmar el estatus somático de aquellas variantes de nueva identificación.

Para tener una visión inicial, se llevó a cabo un control de calidad con *NanoPlot* (De Coster et al., 2018) del archivo de lecturas disponible. El informe detalla la presencia de 18 688 039 lecturas, reducidas a 18 687 976 (63 menos) tras la aplicación de un filtro, con unos datos de longitud y calidad apropiados (Tabla A.2).

A continuación, se llevó a cabo el alineamiento de las lecturas frente al genoma de referencia en su versión GRCh37 con el alineador *Minimap2* (Materiales y métodos 4.2.1). En este punto, se ejecutó un control de calidad con *Qualimap2* (comando *bamqc*) sobre el archivo de alineamientos, cuyas estadísticas principales pueden observarse en Tabla 2.2.

El archivo de alineamientos se empleó para la identificación de indels y SVs con *Sniffles* (Sedlazeck et al., 2018) (Materiales y métodos 4.2.2). Finalmente, se desarrolló el análisis de las variantes obtenidas a punto final del flujo de trabajo (Materiales y métodos 4.2.3). El resultado fue la identificación de 36 390 variantes, cuyos tipos y números correspondientes se detallan en Figura 2.6. De ellas, 25 573 (70.3 %) eran indels y 10 817 (29.7 %) SVs.

	Minimap2 + SAMtools
Tiempo de ejecución (horas)	5.5
Muestra	COLO829
Tamaño de la referencia	3 234 834 689
# de lecturas alineadas	17 614 620 (84.41 %)
# de lecturas no alineadas	3 253 671 (15.59 %)
Longitud de lecturas mín/máx/media	0 / 1 376 732 / 9 717.73
Profundidad de secuenciación media	57.6X

Tabla 2.2: Tiempo de ejecución del alineamiento, ordenamiento y compresión del archivo de formato SAM resultante en el flujo de trabajo de lecturas largas, obtenido mediante la ejecución del comando *seff* en el clúster. A continuación, resumen generado por el control de calidad ejecutado con *Qualimap2*.

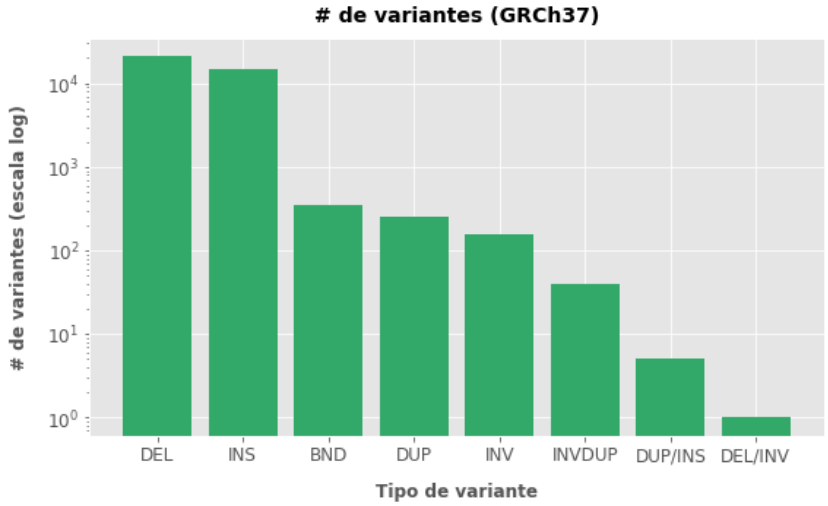

Figura 2.6: Gráfico de barras representando, en escala logarítmica, el número (#) de variantes identificadas por el flujo de trabajo de lecturas largas, para cada tipo de variante.

En este caso, se observaron variantes cuyo campo indicador del cromosoma correspondía a una secuencia de tipo *primary assembly*, pero con la etiqueta *STRANDBIAS*. Este tipo de variantes son las que presentan lecturas que apoyan su presencia en ambas hebras del ADN de forma que, al no poder determinar su posición con exactitud, fueron ignoradas. Por otro lado, se observó la identificación de 355 variantes de tipo BND, que también fueron ignoradas, por las causas expuestas en Resultados 2.1.2. Por último, se filtraron las variantes presentes en los *unlocalized/unplaced scaffolds* del genoma de referencia. Como consecuencia, se obtuvieron 32 229 variantes, entre las que se buscó de forma estricta las 424 variantes de la referencia. Al no encontrar ninguna, se buscó qué variantes de la referencia coincidían únicamente en el cromosoma y la posición en él con las identificadas, ignorando la secuencia, aunque el resultado fue el mismo. Finalmente, se trató de buscar qué variantes de las identificadas se encontraban en un intervalo de ± 500 con respecto a la posición de cada variante de la referencia, obteniendo dos resultados (Tabla A.3).

En el caso de la primera variante encontrada (Tabla A.3; ID 3), la identificada por lecturas largas se trata de una inserción de 109 nucleótidos que se inicia en la posición 110835824 del cromosoma 7. Respecto a la variante de la referencia que coincide con ella, consiste también en una inserción, de 5 nucleótidos de longitud y con una posición de inicio 59 nucleótidos posterior

a la de la identificada por lecturas largas, la cual se encuentra dentro de esta última. Atendiendo a la secuencia de nucleótidos presente en la posición 59 nucleótidos posterior a la de inicio de la inserción de 109 nucleótidos, se observa la secuencia resaltada, prácticamente idéntica a la identificada como insertada en el caso de la variante de la referencia. Hay que indicar que la región en la que se encuentran las variantes descritas presenta una repetición simple.

En este punto, hay que destacar: el criterio de inclusión de una variante en la referencia somática; que la ejecución repetida del flujo de trabajo de lecturas cortas mostró la ausencia de aleatoriedad en el proceso de alineamiento y detección de variantes, al obtenerse siempre los mismos resultados; que el tamaño de la variante propuesta mediante lecturas largas excede el intervalo de longitudes mostrado por las variantes de la referencia y el límite superior impuesto por *Strelka2*; y la participación de las regiones repetitivas del genoma. Estos argumentos y las características de las variantes identificadas por cada tecnología parecen sugerir la posible existencia de errores técnicos debidos a la tecnología empleada en *Illumina* o los programas posteriores a la secuenciación. De tal manera, aunque sí parece haber evidencia de la existencia de una variante determinada, la tecnología propia y asociada a *Illumina* no parece ser capaz de identificarla correctamente, llevando a incluir en la referencia una variante que podría no ser real, la cual es además apoyada por distintas aproximaciones independientes.

En el caso de la segunda variante (Tabla A.3; ID 4), la identificada por lecturas largas se trata de una deleción de 38 nucleótidos, que se inicia en la posición 132332234 del cromosoma 8. La variante de la referencia que coincide con ella consiste en una deleción, de 1 nucleótido de longitud y con una posición de inicio 490 nucleótidos posterior a la de la identificada por lecturas largas. Este caso vuelve a poner de manifiesto la participación de las regiones repetitivas del genoma, al tratarse de variantes ubicadas en una región que incluye una repetición de tipo Alu, una simple (TA) inmediatamente posterior a ella y dos regiones de baja complejidad ricas en AT (Figura 2.7). Así, como se comenta en Introducción 1.1, una lectura larga puede llegar a incluir una parte de la región repetitiva y una no repetitiva previa o posterior a la misma, disponiendo así de un anclaje con el que resolver las repeticiones, como parece ser el caso: aunque *Illumina* permite detectar una deleción de 1 solo nucleótido en la posición 132332724 del cromosoma 8, no es capaz de generar la evidencia suficiente para detectar una de 38 nucleótidos. Del mismo modo, atendiendo a la identificación incluida en la referencia, se demuestra que la detección de variantes con un tamaño inferior a 30 nucleótidos es complicada actualmente con las tecnologías de tercera generación, haciendo todavía necesario el uso de las de segunda generación que, por otro lado, tienen una gran precisión.

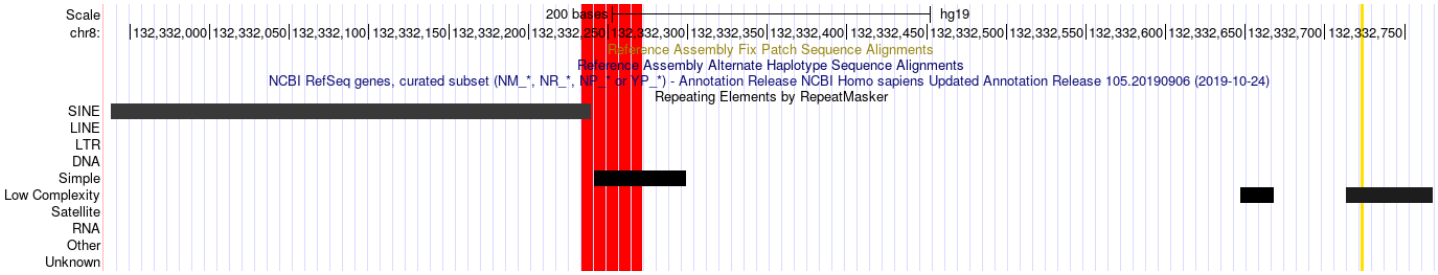

Figura 2.7: Captura del navegador genómico de la UCSC, centrado en una región del cromosoma 8 afectada por una de las variantes estudiadas. La región sombreada en rojo corresponde a la afectada por la variante de interés, presente en el conjunto resultante del flujo de trabajo de lecturas largas, incluyendo las filtraciones descritas. La región sombreada en amarillo corresponde a la afectada por la variante de interés, presente en la referencia somática. El rectángulo gris y los negros indican las regiones correspondientes a elementos genómicos repetitivos.

Volviendo al conjunto total de variantes identificadas por lecturas largas, se realizó un análisis del número de variantes por su longitud. De forma preliminar, llamó la atención la detección de casi 400 variantes con un tamaño cercano a 0, que resultaron ser las 355 variantes de tipo BND,

las cuales volvieron a ser excluidas. El resultado son 36 035 variantes con un tamaño mínimo, medio y máximo de 31, 25 276 y $2.2 \cdot 10^8$ nucleótidos, respectivamente. Además, se observa un pico en el intervalo ± 50 (indels), el cual supone un 40.5 % de las variantes consideradas. Por tanto, se procedió a centrar la visualización del análisis en el intervalo ± 200, para poder observar mejor esta región (Figura 2.8).

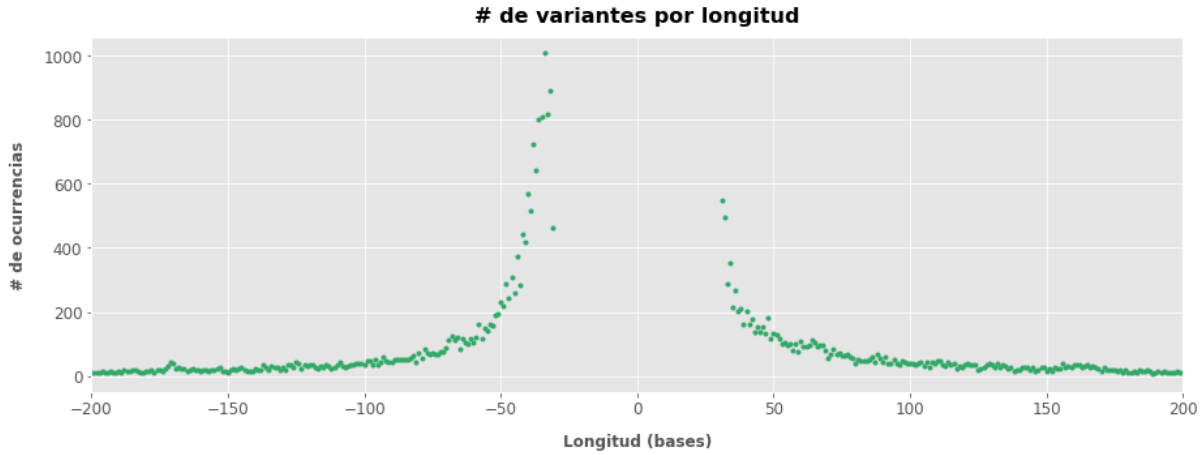

Figura 2.8: Número (#) de variantes por cada valor de longitud posible. Los valores negativos de longitud corresponden a deleciones. El conjunto de variantes empleado corresponde a todas las resultantes del flujo de trabajo de lecturas largas, excluyendo las de tipo BND.

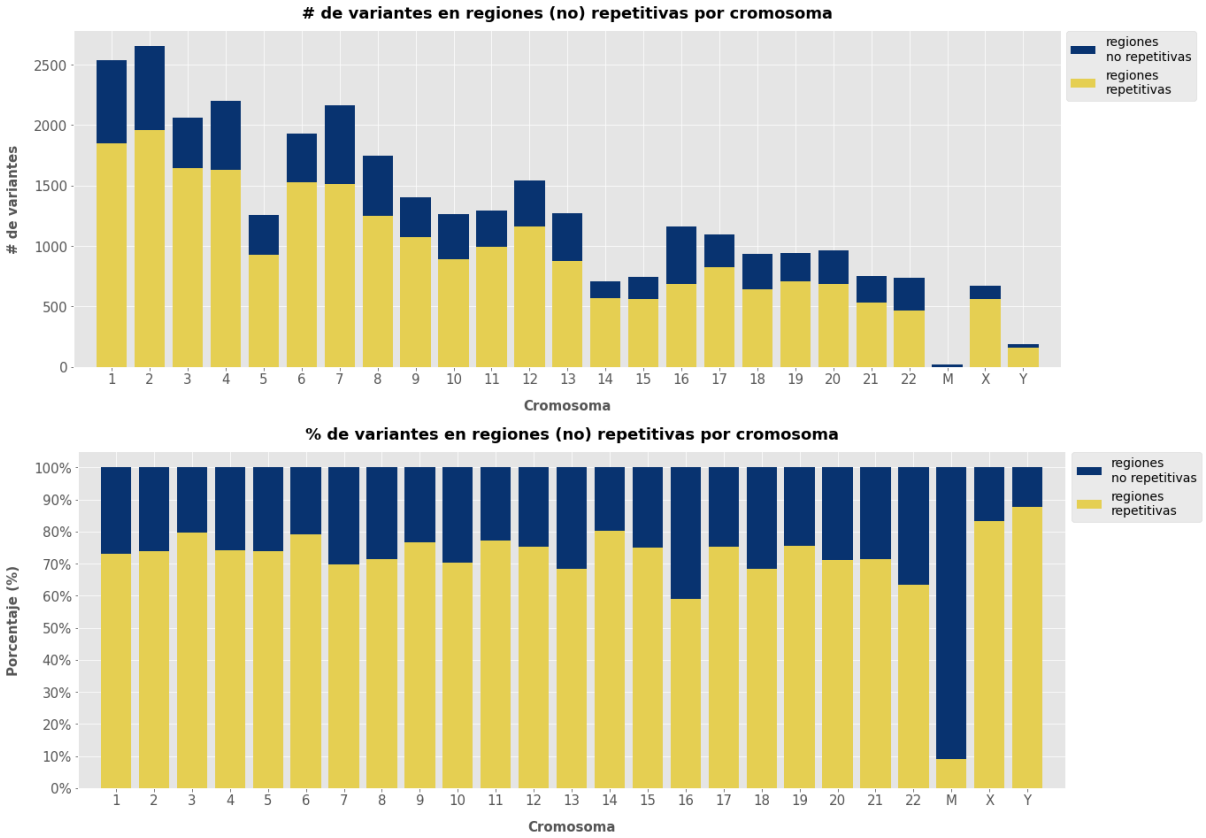

Figura 2.9: Gráficos de barras apiladas representando, para cada cromosoma, el número (#, panel superior) o porcentaje (%, panel inferior) de variantes identificadas por lecturas largas en regiones no repetitivas en azul y el de aquellas en repetitivas en amarillo.

2.2.2. Análisis de la influencia de las regiones repetitivas del genoma

Al igual que en el flujo de trabajo desarrollado para las lecturas cortas, se realizó la búsqueda de las variantes identificadas que caen en regiones repetitivas del genoma. Nuevamente, se excluyeron del proceso todas aquellas variantes y regiones repetitivas que involucran secuencias del tipo *unlocalized/unplaced scaffolds*, así como las variantes etiquetadas como *STRANDBIAS* y las de tipo BND.

Los resultados observados en Figura 2.9 vuelven a poner de manifiesto la gran contribución de las regiones repetitivas en la generación de las indels y SVs identificadas en COLO829. El porcentaje medio de variantes en regiones repetitivas es del 71.29 % (mediana = 73.82 %), siendo el mínimo de variantes en regiones repetitivas del 9.09 % (genoma mitocondrial) y el máximo un 87.63 %. Estos resultados suponen la presencia de un mayor porcentaje de variantes en regiones repetitivas del genoma de COLO829 en el caso de lecturas largas que en el de las cortas. Esto también se cumple en el caso del genoma mitocondrial, en el que la aproximación mediante lecturas cortas no lograba detectar ningún caso. Así, los resultados parecen evidenciar que las lecturas largas permiten estudiar mejor estas regiones, aumentando la resolución en ellas y permitiendo definir mejor qué variantes originan.

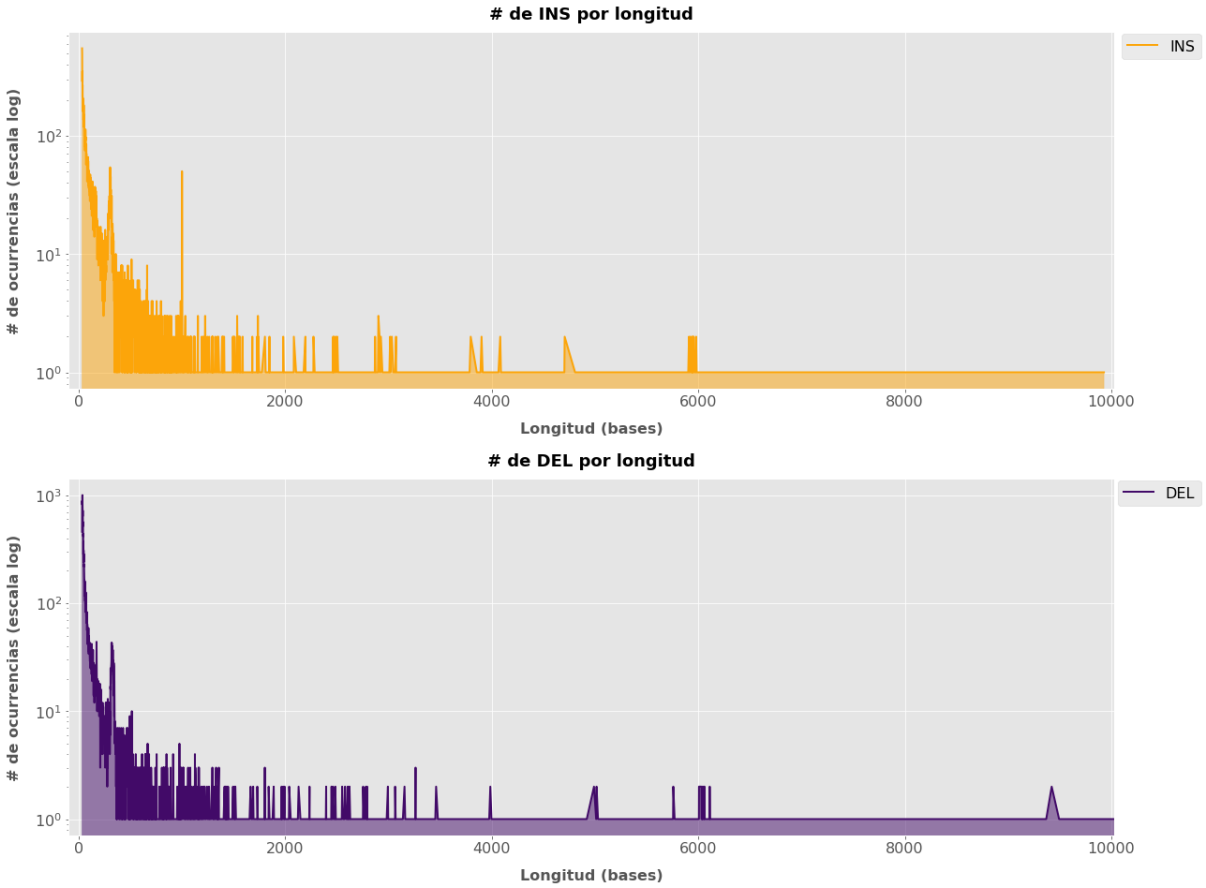

Figura 2.10: Número (#) de variantes, en escala logarítmica, por cada valor longitud posible. En ambos paneles, el conjunto empleado corresponde a todas las variantes resultantes del flujo de trabajo de lecturas largas. En el panel superior, las variantes representadas pertenecen al tipo inserción. En el panel inferior, al tipo deleción, mostrando los tamaños en valor absoluto.

Por otro lado, se estudió el número de variantes por longitud para las variantes de tipo inserción y deleción (Figura 2.10). En ambos casos, llama la atención la presencia de un pico en una longitud en torno a las 300 bases, el cual se ha asociado a elementos repetitivos de tipo Alu, y otro en torno a las 6 000 bases, el cual se ha asociado a elementos repetitivos

de tipo L1 (Nattestad et al., 2018; Ameur et al., 2018; De Coster et al., 2019; Fatima et al., 2020). Para comprobar la consistencia de estos resultados con los obtenidos en este trabajo, se analizaron las inserciones y deleciones presentes en esas longitudes (Figura A.2 y Figura A.3, respectivamente). De 1 644 inserciones con un tamaño comprendido entre 250 y 350 nucleótidos, intervalo en el que se ubica el pico correspondiente, 297 de ellas (18 %) se encuentran en regiones con elementos repetitivos de tipo Alu. De 1 087 deleciones con un tamaño comprendido entre 300 y 350 nucleótidos, 867 de ellas (80 %) se encuentran en esas regiones. De 30 inserciones con un tamaño comprendido entre 5 900 y 6 000 nucleótidos, 24 de ellas (80 %) se encuentran en regiones con elementos repetitivos de tipo L1. De 46 deleciones con un tamaño comprendido entre 6 000 y 6 125 nucleótidos, 44 de ellas (96 %) se encuentran en esas regiones. No sólo estos resultados, sino también la agrupación de la mayoría de las inserciones y deleciones (aproximadamente un 95 % en ambos casos) en un rango de longitud inferior a 1 kilobase, son consistentes con los resultados obtenidos previamente en los trabajos citados. Esto constituye una prueba a favor de la validación de la aproximación seguida en este flujo de trabajo.

2.2.3. Análisis del resto de tipos de SVs

Respecto al resto de tipos de variantes detectados, consiste en un total de 453 variantes de los tipos duplicación (DUP), inversión (INV) y variantes complejas, que son combinaciones de los distintos tipos de SVs (INVDUP, DUP/INS, DEL/INV), de las cuales se calcularon algunas estadísticas (Tabla 2.3).

Tipo de variante	**DUP**	**INV**	**INVDUP**	**DUP/INS**	**DEL/INV**
# de ocurrencias	252	156	39	5	1
Longitud mínima	60	144	258	345	-285
Longitud media	365185.6	4114536.4	1249.4	1000.6	-285
Longitud mediana	3433	14669	655	598	-285
Longitud máxima	75216887	217860447	5591	2638	-285

Tabla 2.3: Número (#) de variantes identificadas por el flujo de trabajo de lecturas largas, para cada tipo de variante compleja. A continuación, estadísticas sobre sus longitudes. Los números negativos corresponden al tamaño de deleciones.

En trabajos anteriores, se ha descrito la limitación del estudio de SVs mediante lecturas cortas, al obtener una tasa de falsos positivos y falsos negativos de hasta el 50 % o superiores (Sudmant et al., 2015; Huddleston et al., 2017). Atendiendo a las estadísticas mostradas en Tabla 2.3, estos resultados son consistentes con los obtenidos en este trabajo pues, como se puede comprobar, hay muchas de las variantes que comprenden tamaños muy superiores a los ofrecidos por lecturas cortas. Así, tratándose de variantes complejas y requiriendo la información aportada por varias lecturas o la aplicación de varias o distintas aproximaciones, su detección resulta comúnmente inalcanzable a través del uso de lecturas cortas.

2.2.4. Análisis de las implicaciones biológicas

Variantes presentes en la referencia somática

En referencia a las implicaciones biológicas que las variantes presentes en COLO829 pudieran tener, en el artículo de la referencia somática (Craig et al., 2016) se resaltan:

- SNV Val600Glu en *BRAF* (protooncogén B-raf, serina / treonina quinasa) que, junto a otras mutaciones, afectan el dominio quinasa y ocurren entre el 30 y 72 % de los melanomas malignos. Además, también se indica la detección, por análisis de CNVs, de una amplificación con dos copias en la región chr7q31.33–36.1.
- Deleción de 2 pares de bases dentro de la secuencia codificante de *CDKN2A* (inhibidor de quinasa dependiente de ciclina 2A).
- Sustitución de las bases de un dinucleótido en el promotor de *TERT* (transcriptasa inversa de la telomerasa). Mutaciones en el promotor *TERT* se han detectado en el 71 % de los melanomas, así como en otras neoplasias malignas.
- Deleción focal de 12 kilobases de *PTEN*, detectada por análisis de CNVs.

Como una aproximación preliminar, se procedió a determinar qué genes se veían afectados por indels y SVs detectadas mediante el flujo de trabajo de lecturas largas. Para ello, se empleó el conjunto de variantes filtrado (excluyendo las presentes en los *unlocalized/unplaced scaffolds*, las etiquetadas como *STRANDBIAS* y las de tipo BND) y el archivo de anotación básica de genes del proyecto *GENCODE* (manteniendo solo los elementos coincidentes con el tipo "gen") (Frankish et al., 2019). El criterio de búsqueda empleado fue el mismo que en el apartado de las regiones repetitivas, con el objetivo de identificar no solo las variantes contenidas en el cuerpo del gen, sino también las que pudieran afectar a su parte inicial o final sin estar completamente dentro del mismo (Figura 2.4). El resultado fue la identificación de 11 613 genes únicos alterados, excluyendo pseudogenes.

Centrando la atención en algunos de los genes del análisis de referencia resaltados, si bien la deleción de 2 pares de bases de *CDKN2A* no se puede detectar con la aproximación basada en lecturas largas, llama la atención una gran deleción que afecta al cromosoma 9, incluyendo el citado gen, la cual abarca casi 24 megabases (chr9, 6727584 - 30688960). En el caso de *TERT*, se detecta una deleción de 53 nucleótidos y tres inserciones de 62, 101 y 80 nucleótidos, no solapantes entre ellas.

Sin embargo, la identificación más interesante lograda en el presente trabajo es la de una variante en *PTEN*, que coincide con lo reportado en la literatura, pero logra dicha identificación de manera directa y más precisa. En la construcción preliminar de la referencia somática de COLO829 (Pleasance et al., 2010), el análisis de CNVs detecta una deleción homocigótica interna de 12 kilobases en *PTEN* (chr10: 89690279 - 89702321), de la que se predice un efecto de terminación prematura con probable implicación en el desarrollo de COLO829, al tratarse de un gen supresor de tumores. Posteriormente (Craig et al., 2016), otro análisis de CNVs vuelve a detectar una deleción de tamaño similar, de la que se predice un efecto de pérdida de función. Además, se indica que dicha variante se haya reportada en *COSMIC* para COLO829. La realidad es que la única disminución en el número de copias reportada en *COSMIC* tiene un tamaño de 9 513 bases (chr10: 89700671 - 89710183) (Figura 2.11).

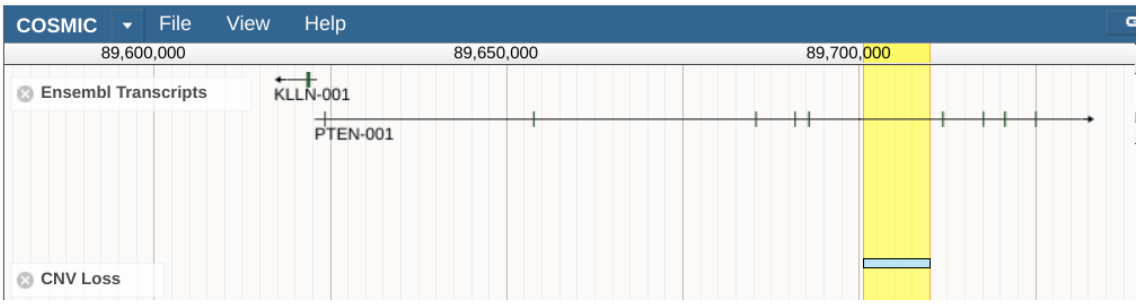

Figura 2.11: Captura del navegador genómico de *COSMIC*, centrado en una región del cromosoma 10 afectada por una de las variantes estudiadas. La región sombreada en amarillo corresponde a la afectada por la variante de interés.

En la detección de variantes mediante CNV, la secuenciación de genoma completo permite establecer una línea de base, es decir, estimar el número normal de lecturas que debería haber, y así poder detectar alteraciones del mismo. De esta manera, el análisis de CNVs se basa en la agrupación de las lecturas que hay en una ventana de cierta longitud del genoma, con las que se trata de identificar un conjunto común que las relacione, del cual se infiere si está en una cantidad apropiada o, por el contrario, superior (ganancia) o inferior (pérdida) a la esperada. Sin embargo, esta aproximación es indirecta y de una resolución limitada.

A diferencia de esta aproximación indirecta, en el caso de la metodología implementada en este trabajo, la determinación de variantes somáticas con lecturas largas identifica directamente la variante implicada, conociendo incluso la secuencia afectada y su longitud con exactitud (chr10: 89700300 - 89712341, 12 041 bases) (Figura 2.12). Este ejemplo evidencia el gran potencial del uso de lecturas largas en el estudio de SVs y supone una validación de la aproximación desarrollada en este trabajo, al tratarse de una variante ya conocida. En contraposición, una sola o varias lecturas cortas no suelen abarcar el considerable rango de tamaños posibles de las SVs, al definirse como variantes de al menos 50 nucleótidos, lo que supone así un mayor requerimiento para poder identificarlas. Aunque, en este caso particular, la limitación se consigue salvar a través del análisis de CNVs, obteniendo un resultado aproximado, no suele ocurrir en todos los casos.

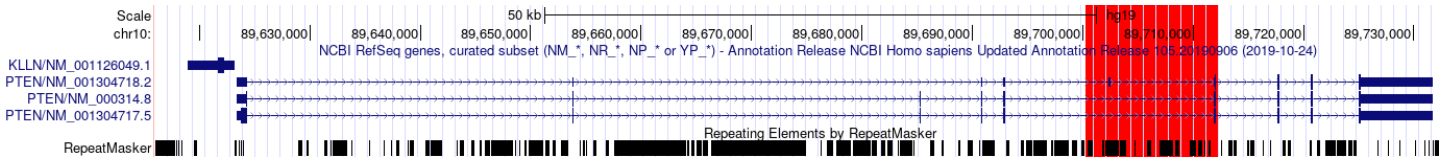

Figura 2.12: Captura del navegador genómico de la UCSC, centrado en una región del cromosoma 10 afectada por una de las variantes estudiadas. La región sombreada en rojo corresponde a la afectada por la variante de interés, presente en el conjunto resultante del flujo de trabajo de lecturas largas, incluyendo las filtraciones descritas. Los rectángulos negros indican las regiones correspondienteS a elementos genómicos repetitivos.

Detección de nuevas variantes de interés

Habiendo analizado la capacidad de la aproximación presentada para detectar variantes conocidas, se procedió a estudiar su capacidad para proponer nuevas variantes. A través de *TumorPortal* (`http://www.tumorportal.org/`) se tomó la lista de genes cuyo impacto está clasificado en tres categorías: altamente significativo, significativo, o asociado en el desarrollo de melanomas. Entre los altamente significativos, se encuentra *CDKN2A*, del que se han detectado SVs, como ya se ha descrito. Otro gen altamente significativo es *TP53*, del cual no se detectan alteraciones en un 90 % de los melanomas del subtipo definido por las mutaciones en *BRAF* (Akbani et al., 2015). En nuestro caso, tampoco se han encontrado variantes en *TP53*, pero sí una deleción de casi 52.5 kilobases en una región del cromosoma 16 que contiene al gen *TP53TG3B*, sobre el que existe evidencia de su participación en la ruta de *TP53* (Ng et al., 1999). Así, aunque no haya alteraciones en *TP53*, alteraciones de su ruta podrían tener también implicaciones importantes, por ejemplo, la pérdida de una de sus vías de supresión del desarrollo tumoral. De esta forma, el uso de lecturas largas permite acceder a una fuente de información antes desconocida, de la que poder conocer más aspectos relativos al tumor de estudio.

También se buscaron cuántos de los 11 613 genes únicos alterados forman parte de la lista de 576 genes presentes en el *COSMIC Cancer Gene Census*, los cuales consisten en genes mutados con implicaciones en el desarrollo del cáncer (Sondka et al., 2018). Como resultado, se encontró que 558 variantes de las identificadas afectan a 243 genes únicos (42 %) implicados en el desarrollo del cáncer, de las cuales 316 (56.6 %) son SVs y 242 (43.4 %) son indels. Además, ninguna de las

variantes encontradas se encuentra dentro del rango de tamaños de las variantes constituyentes de la referencia somática de COLO829.

Por último, como se puede comprobar en la Tabla 2.3, llamó la atención la existencia de una sola variante de tipo DEL/INV, cuyo tamaño es de 285 nucleótidos. Al buscar si dicha variante afectaba a algún gen en particular, se encontró que afectaba al gen *RPH3AL*, un gen con alta probabilidad de tener un papel como supresor de tumores, el cual se ha estudiado en relación con el cáncer de pecho (Putcha et al., 2015). A día de hoy, se desconoce mucha información sobre la metástasis, a pesar de su importancia en el cáncer y sus consecuencias para la salud de los pacientes (constituye la principal causa de muerte para más del 90 % de ellos). Sin embargo, se ha visto que el microambiente de los tejidos de desarrollo de tumores secundarios son preparados por el tumor primario incluso antes de que se produzca la metástasis. En este proceso, se ha observado la participación de factores secretores y vesículas extracelulares que, entre otros componentes, transportan ARNm de genes que participan en la metástasis los cuales, al ser recibidos por células sanas, parecen provocar el desarrollo de un fenotipo maligno (Fares et al., 2020). Del mismo modo, la detección de variantes en genes implicados en el desarrollo de tipos de tumores distintos al primario plantea la posibilidad de que éstas puedan tener también algún tipo de implicación en el desarrollo de los tumores secundarios.

2.3. Comparación de los flujos de trabajo desarrollados

Inevitablemente, el análisis independiente de los resultados de cada flujo de trabajo presentado en las secciones anteriores incluye algunas comparaciones entre ellos. Sin embargo, como parte final del análisis, se ha procedido a realizar una comparación más explícita. En primer lugar, hay que destacar que el conjunto empleado de variantes identificadas por lecturas cortas ha sido: el generado por *Strelka2* y filtrado por las secuencias en las que se detectaron (se excluyen variantes que involucran a *unlocalized/unplaced scaffolds*), formado por 128 919 variantes; el mismo, pero con el filtro adicional de acuerdo al campo *FILTER* del archivo de formato VCF, formado por 1 411 variantes; y el generado por *Manta* y filtrado por las secuencias en las que se detectaron (se excluyen variantes que involucran a *unlocalized/unplaced scaffolds*), formado por 141 890 variantes. Respecto al conjunto de variantes identificadas por lecturas largas, se ha empleado el filtrado, excluyendo las presentes en los *unlocalized/unplaced scaffolds*, las etiquetadas como *STRANDBIAS* y las de tipo BND, el cual consiste en 32 229 variantes.

Una vez definidos los conjuntos de variantes a considerar, se procedió a buscar, por cada variante del conjunto originado por lecturas largas, aquellas del conjunto originado por lecturas cortas coincidentes en el mismo cromosoma y misma posición en él. El mismo procedimiento se aplicó para los tres conjuntos de variantes originados por lecturas cortas, resultando en la identificación de 24, 2 y 598 variantes, respectivamente, no correspondiendo ninguna a variantes presentes en la referencia somática de COLO829. Hay que destacar que, en el caso de la búsqueda entre las variantes identificadas por *Manta*, se consiguió simplificar a 527 variantes, al no incluir las definidas como de tipo BND.

Respecto a los resultados obtenidos entre las variantes identificadas por lecturas largas y las identificadas por cortas con *Strelka2*, se va a centrar la atención en las dos variantes resultantes al usar el conjunto más filtrado, al sintetizar de la forma más eficiente las observaciones realizadas (Tabla A.4; ID 5, ID 6).

En el caso de la primera (Tabla A.4; ID 5), se trata de una pequeña deleción (< 50 nucleótidos) identificada en la posición 118516186 del cromosoma 1 por ambas tecnologías. Sin embargo, la secuencia identificada como presente en la referencia y, por tanto, aquella que se deleciona, no coincide en su totalidad entre ambas estrategias (la porción de secuencia coincidente se resalta-da en negrita). Quizás, la diferencia entre ambas tecnologías se explique por la mayor tasa de

error de las plataformas de *Oxford Nanopore Technologies* (Logsdon et al., 2020), cuyo efecto se amplifica a menor tamaño de la variante, aunque no impide su identificación. Otra posibilidad es la influencia de regiones repetitivas porque, si bien no hay ninguna en el intervalo de posiciones afectado por la variante, se ubica en una región muy próxima al centrómero, con alta presencia y actividad de este tipo de elementos genómicos.

En el caso de la segunda (Tabla A.4; ID 6), se trata de una inserción (> 50 nucleótidos) identificada en la posición 60398423 del cromosoma 12 por ambas tecnologías. Sin embargo, la secuencia identificada como constituyente de la variante difiere entre ambas estrategias. En el caso de *Illumina*, se describe la inserción de una adenina mientras que, en el de ONT, se trata de una inserción de 126 nucleótidos. Si bien *Illumina* parece tener evidencias de que algo ocurre en esa región, la cual presenta una repetición simple del tipo (TA)n (Figura 2.13), no es capaz de definir con exactitud la variación real con respecto a la referencia. Sin embargo, la variante identificada por lecturas largas, empezando en la región repetida, se extiende más allá de su final, solucionando la limitación de las lecturas cortas.

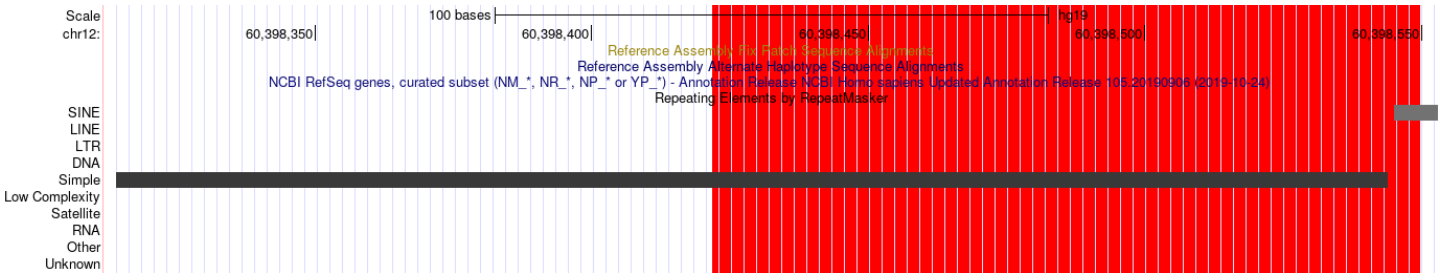

Figura 2.13: Captura del navegador genómico de la UCSC, centrado en una región del cromosoma 12 afectada por una de las variantes estudiadas. La región sombreada en rojo corresponde a la afectada por la variante de interés presente en el conjunto resultante del flujo de trabajo de lecturas largas, incluyendo las filtraciones descritas. La región correspondiente a la afectada por la variante de interés presente en el conjunto resultante del flujo de trabajo de lecturas cortas se haya oculta por suponer un tamaño (1 nucleótido) muy inferior al de la anterior, aunque se ubica exactamente en el inicio de la región sombreada. Los rectángulos grises indican las regiones correspondientes a elementos genómicos repetitivos.

Respecto a los resultados obtenidos entre las variantes identificadas por lecturas largas y las identificadas por cortas con *Manta*, centrar la atención en 527 variantes supone una complicación que se ha tratado de salvar buscando cuáles de ellas forman parte también del conjunto identificado por *Strelka2*, resultando en un total de 10, de las que solo 1 se encuentra en el conjunto más filtrado. Esta última resulta ser la variante de la posición 118516186 del cromosoma 1 recientemente descrita, de lo que se infiere que consiste en una de esas variantes candidatas ofrecidas por *Manta* a *Strelka2* que este último incluye en sus resultados finales. De entre las demás variantes, llama la atención la deleción de la posición 69687148 del cromosoma 10, cuya secuencia identificada como presente en la referencia y, por tanto, aquella que se deleciona, no coincide en su totalidad entre ambas estrategias (la porción de secuencia coincidente se resalta en negrita), habiendo solo 3 nucleótidos de diferencia (Tabla A.4; ID 7). La variante se encuentra en una región con una repetición simple, cuyas características no parecen dificultar su identificación incluso con lecturas cortas. Esta observación también podría sugerir que las limitaciones de *Illumina* en regiones repetitivas no se deben siempre a la longitud de las lecturas, sino también a problemas técnicos asociados a los programas de identificación de variantes.

No obstante, también hay otros ejemplos, como el de la variante de la posición 72217526 del cromosoma 15, también ubicada en una repetición simple. En el caso de los resultados obtenidos con lecturas largas o con cortas con *Manta*, esta variante consiste en una deleción cuya secuencia identificada como presente en la referencia por ambas estrategias coincide en una gran porción (resaltada en negrita) (Tabla A.4; ID 8). Teniendo esas variantes una longitud de 52 y 50 nucleótidos, respectivamente, *Strelka2* reporta en la misma posición una deleción de

tan solo 2 nucleótidos. Este resultado parece volver a sugerir la existencia de errores técnicos en los programas de identificación de variantes ya que, si bien la variante identificada por lecturas largas y por cortas con *Manta* son muy parecidas, nada tienen que ver con la identificada por cortas con *Strelka2*, el cual usa las mismas lecturas que *Manta*.

Hay que recordar que la búsqueda de las 424 variantes de la referencia somática entre las identificadas por lecturas largas, usando un intervalo de la posición de cada variante de la referencia considerada ± 500, llevó a la identificación de dos variantes. Inicialmente, éstas no se encontraron entre las variantes coincidentes de cualquiera de las búsquedas realizadas en este apartado (Resultados 2.3). Sin embargo, al buscarlas directamente en el conjunto de variantes identificadas por lecturas cortas con *Strelka2*, usando el mismo intervalo de posición que en la búsqueda original, se obtuvieron las mismas variantes de la referencia. Y, al repetir la búsqueda en el conjunto de variantes identificadas por lecturas cortas con *Manta*, se obtuvo: para la variante del cromosoma 7, la misma variante identificada por lecturas largas (en amarillo la hipotética secuencia coincidente entre las identificadas por *Manta* y lecturas largas con la de la referencia); y, para la variante del cromosoma 8, una variante del mismo tipo (deleción) que la identificada por lecturas largas, cuya posición de inicio es 34 nucleótidos posterior a la misma (en negrita la porción de secuencia coincidente) (Tabla A.4; ID 9, ID 10). Estos resultados refuerzan las conclusiones alcanzadas previamente.

Las 24 variantes coincidentes entre las variantes identificadas por lecturas largas y las identificadas por lecturas cortas resultaron afectar a 18 genes del conjunto de anotaciones básicas de genes del proyecto *GENCODE*, de los cuales ninguno aparece en el conjunto de genes del *COSMIC Cancer Gene Census*. Respecto a las 527 variantes coincidentes entre las variantes identificadas por lecturas largas y las identificadas por lecturas cortas con *Manta*, resultaron afectar a 254 genes del conjunto de anotaciones básicas de genes del proyecto *GENCODE*, de los cuales 8 aparecen en el conjunto de genes del *COSMIC Cancer Gene Census*: *BMPR1A*, *CREB3L1*, *NCOR1*, *XPO1*, *LRP1B*, *PIK3CB*, *MAP3K13*, *AFF1*. En ninguna de las publicaciones centradas en el estudio de las variantes y genes implicados en el desarrollo de COLO829 se hace referencia a ninguno de estos 8 genes destacados (Pleasance et al., 2010; Akbani et al., 2015; Craig et al., 2016).

3

Conclusiones y trabajo futuro

El cáncer es una enfermedad compleja debida a alteraciones genéticas o epigenéticas en el ADN, cuyos cambios pueden ser germinales o somáticos y contribuyen a desarrollar la patología (Stratton et al., 2009). En su estudio, la aparición y el desarrollo de las técnicas de secuenciación masiva ha sido clave y, en particular, la secuenciación de lecturas cortas y tipo *paired-end* correspondiente a *Illumina* (Illumina, Inc., 2021). Por el contrario, su propia naturaleza también ha limitado el ensamblaje del genoma humano (Figura 1.2) y la detección de muchas variantes y tipos de ellas, como las SVs (Sudmant et al., 2015; Logsdon et al., 2020). Por ello, otras aproximaciones alternativas, basadas en lecturas cortas y largas, surgieron para solventar los obstáculos enfrentados. Entre ellas, destaca la secuenciación basada en nanoporos de ONT que, en lugar de usar la incorporación o hibridación de nucleótidos a una hebra de ADN molde, detectan directamente la composición de dicha hebra (Figura 1.1b) (Oxford Nanopore Technologies, 2020). A pesar del reciente establecimiento de las aproximaciones basadas en lecturas largas en la investigación y la clínica, su uso ha aumentado el conocimiento de distintos tipos de variantes genéticas humanas, sobre todo SVs. No obstante, también se han encontrado problemas y limitaciones que hacen seguir necesitando tanto la secuenciación de segunda generación, como la optimización de los programas ya disponibles y el desarrollo de otros nuevos para la de tercera generación (Chaisson et al., 2019; Logsdon et al., 2020; Spealman et al., 2020).

Por tanto, el objetivo de este trabajo es el de definir capacidades y limitaciones concretas de las tecnologías de secuenciación de segunda y tercera generación en la detección y el estudio de SVs en cáncer. Esto no sólo permitiría aumentar el conocimiento sobre el tipo de cáncer en sí, sino también sobre otros tipos de cáncer, la metástasis, el desarrollo de aproximaciones terapéuticas, etc.

Para alcanzar este objetivo, se ha trabajado sobre datos de secuenciación de genoma completo de las líneas celulares COLO829 (melanoma cutáneo metastático) y COLO829BL (línea linfoblastoide sana del mismo paciente) (Morse and Moore, 1993). Respecto al melanoma, según el programa de Vigilancia, Epidemiología y Resultados Finales (SEER) del Instituto Nacional del Cáncer de los Estados Unidos de América, sus tasas han ido en aumento en las últimas décadas, llegando a constituir el quinto tipo de cáncer más común. Su origen suele encontrarse en la exposición prolongada a la radiación ultravioleta de la luz solar natural o artificial, la cual provoca un daño acumulativo. En un 83 % de los casos, el diagnóstico se produce estando la enfermedad localizada (confinada al sitio primario). En un 9 %, en el estado de diseminación a los ganglios linfáticos regionales. En un 4 %, en el estado de metástasis distal. Y en el último 4 %, se diagnostica sin conocer el estado. Según el programa SEER, las personas con enfermedad

metastásica distante, estado más grave, tienen una tasa de supervivencia a 5 años del 27.3 % (Sundararajan et al., 2020). Así, dado el impacto de este tipo de cáncer, y disponiendo de datos de secuenciación de lecturas cortas y largas con alta profundidad, se decidió usar como base de este trabajo.

En el caso de las lecturas cortas, la disponibilidad de una referencia somática para COLO829 (Craig et al., 2016) se usó como punto de partida. De esta forma, la identificación de indels somáticas desarrollada en este trabajo, usando un solo conjunto de lecturas (de cuatro totales) obtenidas mediante secuenciación de tipo *paired-end* con *Illumina* y dos pasos de filtración, permitió recuperar más de un 90 % de las variantes presentes en la referencia, usando solo un 1.1 % de las variantes identificadas (128 919). Así, se dispone de una validación del flujo de trabajo creado, incluyendo las filtraciones, de lo que se infiere la eliminación de un gran número de falsos positivos y se aporta un conjunto de variantes que pueden ser estudiadas para comprobar su existencia y significancia. A pesar de ello, el resto de resultados asociados también sugieren la presencia de falsos positivos no filtrados en un rango de longitudes de ± 4, así como la posible existencia de falsos negativos fuera del rango (-25, 20) (Figura 2.1, panel inferior).

Por otro lado, la repetición de este flujo de trabajo sustituyendo el alineador empleado por *Minimap2* consiguió prácticamente los mismos resultados, perdiendo una sola variante y suponiendo una reducción del tiempo de ejecución en el paso de alineamiento de más de la mitad (Tabla 2.1). Esto supone una validación adicional del flujo de trabajo creado y apoya el reemplazamiento de *BWA* por *Minimap2* en el paso de alineamiento de flujos de trabajo para la identificación de indels y SVs, con su consiguiente reducción en el tiempo de ejecución requerido.

En un intento preliminar de proponer variantes excluidas de la referencia somática, se procedió a usar el conjunto de variantes identificado con *Manta*, en el que se encontraron 6 variantes coincidentes en el cromosoma y posición en él (al menos, aproximada), pero no en cuanto a la secuencia constituyente de la variante. Dada la presencia de 5 de ellas en regiones repetitivas del genoma y la contribución de estas regiones en la generación del tipo de variantes estudiadas (un 64.3 % de las identificadas por *Strelka2*), estos resultados evidencian la limitación de las lecturas cortas en tales regiones y sugieren que la variante real podría ser distinta a la reportada.

En el caso de las lecturas largas, la aplicación de filtros equivalentes y uno adicional respecto a las variantes de tipo BND, resultó en la identificación de 32 229 variantes, manteniéndose así más de un 88 % de las totales, en contraposición al 1 % mantenido con las lecturas cortas. De las variantes consideradas, dos de ellas mostraban relación con la referencia somática, además de ser también identificadas en el caso de *Strelka2* y *Manta* (Tabla A.4; ID 9, ID 10). Su estudio en profundidad refuerza la posible existencia de errores técnicos asociados a lecturas cortas, al incluirse en la referencia variantes que parecen ser distintas de las reales, y demuestra su incapacidad en regiones repetitivas del genoma, así como la necesidad de su aplicación para variantes menores de 30 nucleótidos.

El resto de comparaciones realizadas, enfrentando las variantes identificadas por lecturas largas frente a las identificadas por *Strelka2* y *Manta*, permitió centrar la atención en 24 y 527 variantes respectivamente coincidentes. El análisis profundo de algunas de ellas evidencia el papel de las regiones repetitivas del genoma en la generación de variantes de tipo indel y estructurales (Tabla A.4). Si bien las tecnologías de secuenciación de tercera generación demuestran un mejor rendimiento en estas regiones, su mayor tasa de error (Logsdon et al., 2020) determina su papel como mejora y complementación de las de segunda generación, pero no su reemplazamiento. De forma adicional, parece haber un sesgo en la tecnología propia o posterior a la secuenciación por lecturas cortas que, si bien no afecta a *Manta* en algunos casos, sí afecta a *Strelka2*, llevando a la exclusión de variantes de la referencia somática disponible para COLO829 (Craig et al., 2016). Por consiguiente, la aplicación simultánea de flujos de trabajo sobre lecturas cortas y largas, como se desarrolla en este trabajo, demuestra ser una estrategia para plantear más

variantes candidatas que, aunque pendientes de ser validadas experimentalmente y comprobarse su significancia, permiten ampliar el catálogo de variantes ya identificadas.

Respecto al posible impacto de las variantes desconocidas e identificadas por lecturas largas, afectan a 11 613 genes únicos, entre los genes aportados por *GENCODE*. Además, 558 de las variantes responsables de esas alteraciones (56.6 %, SVs; 43.4 %, indels) afectan a 243 genes implicados en el desarrollo del cáncer (42 % de los descritos). Con ello, se aumenta el conocimiento sobre variantes de significancia clínica ya descritas (deleción de 12 kilobases en *PTEN*, Figura 2.12), al identificarse con mayor resolución y fiabilidad, de la misma manera que sobre otras no conocidas, cuyo descubrimiento permite conocer mejor el caso de estudio o, al menos, plantear posibles rutas involucradas en el desarrollo tumoral (*TP53TG3B*), incluyendo el desarrollo de la metástasis (*RPH3AL*).

Insistiendo en las ventajas aportadas por las lecturas largas, este trabajo no sólo ha contribuido a demostrar y reforzar su papel clave en regiones repetitivas, sino también de cara a la detección de variantes complejas, la cual no es posible a través de lecturas cortas.

De cara al futuro, centrándose de forma específica en el flujo de trabajo desarrollado, se debe incluir el análisis de las BNDs, ya que constituyen una fuente de variantes complejas que no han podido ser comparadas entre las dos aproximaciones aplicadas. Sin embargo, no hay que olvidar su relación con las regiones repetitivas, resultando en su no identificación al clasificarse como otros tipos y en un incremento de falsos positivos, como se ha descrito previamente (Abel et al., 2020; Gong et al., 2020).

De forma equivalente, tampoco hay que obviar que las limitaciones del análisis también se deben a la necesidad de una referencia que, al haber sido construida con lecturas cortas, se ven afectadas por su naturaleza intrínseca. En lo respectivo a este trabajo, se deberían actualizar los resultados a la nueva versión del genoma de referencia, para poder así beneficiarse de su información corregida o adicional. También se plantea la construcción de nuevas referencias u otras basadas en aproximaciones distintas, para poder recoger toda la variabilidad genética y así poder aplicarlas de forma más específica a cada población (Yang et al., 2019).

Los problemas asociados a la identificación de falsos positivos, predominantes entre las variantes identificadas por lecturas cortas, también deberían ser tratados. Para ello, se pretende incluir otro programa de identificación de indels y SVs con lecturas cortas, de forma que la intersección de sus resultados con los obtenidos por *Manta* sirva de filtro adicional. Entre los candidatos posibles, *Delly* (Rausch et al., 2012) se muestra como una opción de peso, dado su rendimiento, la automatización de la identificación de variantes somáticas similar *Manta* (Gong et al., 2020) y sus recientes actualizaciones. Entre ellas, resulta interesante el análisis de CNVs, que podría incorporarse. Esto plantea mayores problemas en el caso de las lecturas largas, en el que faltan todavía hacen falta programas estándar de análisis de dichas variantes. En este trabajo, se ha tratado de ejecutar *Nano-GLADIATOR* por todos los medios (Magi et al., 2019). No obstante, el resultado ha sido decepcionante, en gran parte provocado por una documentación y código asociados confusos, y prácticamente imposibles de ejecutar por cualquier usuario distinto del creador.

También es necesario el desarrollo de programas estándar de anotación de SVs. En sí mismo, esto requiere un gran trabajo de cara a aplicar masivamente la secuenciación de tercera generación para la identificación y validación de este tipo de variantes, permitiendo así un incremento de su presencia en las bases de datos, que ahora es muy inferior al que debería. Dado el predominante peso de las variantes somáticas en el desarrollo del cáncer, este trabajo es esencial en su estudio, así como en el de otras patologías con base genética. En otras palabras, se requiere la construcción de referencias somáticas para distintos tipos de cáncer, con las cuales poder evaluar y forzar la optimización de las aproximaciones desarrolladas. Hasta ahora, esto se ha hecho sobre todo con datos simulados que, si bien no son ni mucho menos prescindibles, también se requiere

de casos reales. De esta forma, también sería interesante la aplicación de los flujos de trabajo desarrollados en este proyecto sobre datos sintéticos, para tratar de evaluar cuál podría ser su rendimiento en otros contextos. Además, en relación con todo esto, se debería llevar a cabo la secuenciación con ONT de la línea celular normal asociada al cáncer analizado en este trabajo (COLO829BL), para confirmar aquellas variantes somáticas de nueva identificación.

En conclusión, el desarrollo tecnológico y la investigación científica han permitido aumentar nuestro conocimiento del genoma humano, su funcionamiento y sus implicaciones sobre la evolución, la vida y la salud humana. Recientemente, las mejoras de precisión, escalabilidad y coste de las tecnologías de secuenciación de tercera generación han permitido su establecimiento como una alternativa real y útil, y la consecuente extensión en su uso. Esto conlleva avances especialmente relevantes para el estudio de enfermedades complejas como el cáncer, dada la participación esencial en su desarrollo de variantes genéticas como las estructurales, cuya identificación se ha visto limitada. Al mismo tiempo, esto deriva en la identificación de problemas y la falta de algunos medios para analizar los datos generados. A pesar de ello, su planteamiento permite el desarrollo de soluciones, a veces con diferente enfoque y resultados, muchos de los cuales útiles y complementarios.

Por tanto, el futuro (casi ya presente) de la secuenciación genética va a ser fundamental para el desarrollo definitivo de la medicina personalizada o de precisión, cuyo papel en el campo de la oncología es clave para la prevención del cáncer, la detección del mismo y su posible recurrencia, y para la elección o desarrollo del tratamiento y la predicción de la respuesta del paciente ante el mismo. Esto se debe a que tanto el individuo estudiado como las circunstancias en las que se desarrolla son únicas, pudiendo darse características comunes entre pacientes del mismo tipo de cáncer o incluso de distintos tipos de cáncer, del mismo modo que características específicas de cada tipo de cáncer o paciente, que serían así abordadas de forma específica.

4

Materiales y métodos

4.1. Flujo de trabajo basado en secuenciación de segunda generación

4.1.1. Creación de los archivos en formato FASTQ

Sobre los archivos de alineamiento en formato BAM generados por GSC (número de acceso del *European Genome-phenome Archive*, EGAS00001001385), se aplicó el programa *SAMtools* (comandos, *sort* y *fastq*; versión 1.11) (Li et al., 2009), extrayendo así los archivos de lecturas para cada línea celular (COLO829, COLO829BL) en formato FASTQ. Cabe destacar que la secuenciación realizada era de genoma completo y tipo *paired-end* de forma que, para cada línea celular, se creó un archivo de lecturas conteniendo aquellas obtenidas desde un extremo (*forward*) y otro con las obtenidas desde el otro (*reverse*), pudiendo así partir de los datos inmediatamente generados por la plataforma de secuenciación.

4.1.2. Alineamiento

Las lecturas *paired-end* de cada línea celular (125 nucleótidos) se alinearon frente al genoma humano de referencia en su versión GRCh37 (https://www.ncbi.nlm.nih.gov/genome/guide/human/), con el alineador *BWA* (comando *mem*; versión 0.7.17) (Li, 2013) o *Minimap2* (versión 2.17) (Li, 2018), dando como resultado un archivo con los alineamientos en formato SAM (*Sequence Alignment Map*), los cuales fueron ordenados y almacenados en un archivo comprimido (formato BAM) por el programa *SAMtools* (comando *sort*; versión 1.11).

4.1.3. Marcaje y exclusión de lecturas duplicadas

Las lecturas duplicadas en los archivos de alineamiento en formato BAM de cada línea celular fueron marcadas y eliminadas mediante el programa *Sambamba* (comando *markdup*, opción *-r*; versión 0.7.1) (Tarasov et al., 2015).

4.1.4. Identificación de variantes somáticas

Se ejecutó *Strelka2* (Kim et al., 2018) sobre los archivos de alineamiento en formato BAM de cada línea celular, siguiendo las recomendaciones esgrimidas en la configuración somática del manual de uso disponible en `https://github.com/Illumina/strelka`, incluyendo la aportación de un archivo de formato VCF con variantes candidatas, obtenido a través de la ejecución previa de *Manta* (Chen et al., 2016), siguiendo las mismas recomendaciones del manual de uso disponible en `https://github.com/Illumina/manta`. *Manta* es un programa que puede emplearse para la identificación de variantes somáticas de tipo indel y estructurales, el cual ofrece distintos archivos de salida en formato VCF, de los cuales se aportó aquel con menor limitación (*candidateSV.vcf.gz*), por contener todas las variantes identificadas y no incluir ningún tipo de *ranking* para las mismas. Sin embargo, las SVs candidatas son ignoradas por *Strelka2*, al presentar un límite de longitud máximo para las variantes identificadas de 49 nucleótidos, pudiendo sólo identificar SNVs e indels.

4.1.5. Análisis de las variantes identificadas

Las diferentes estadísticas, búsquedas, tablas y figuras correspondientes a análisis realizados sobre las variantes identificadas se obtuvieron con código escrito en *Python* (Van Rossum and Drake, 2009). Para ello, se hizo uso de *Jupyter Notebook* (Kluyver et al., 2016), aplicación que permite mezclar texto, código (y su ejecución) y visualizaciones. En este ambiente, se han usado módulos de *Python* como: *NumPy* (Harris et al., 2020), cuyas operaciones de vectorización, indexación y transmisión permiten llevar a cabo computaciones numéricas de forma rápida y versátil; *pandas* (Jeff Reback et al., 2021), que permite importar, crear y manejar datos de forma rápida y eficiente gracias a objetos del tipo *DataFrame* y operaciones vectoriales; o *Matplotlib* (Hunter, 2007), que permite crear distintas visualizaciones con los datos resultantes.

Respecto al proceso de búsqueda de intersecciones entre variantes y genes o regiones repetitivas del genoma, hay que destacar que consistió en: por cada variante, se seleccionaron todas las regiones repetitivas presentes en el mismo cromosoma que la misma y, de ellas, se comprobó si alguna presentaba una intersección de al menos un nucleótido. Esto permite la identificación tanto de las variantes ubicadas dentro de la región considerada, como la de aquellas que solo afectan a la parte inicial o final de las mismas sin encontrarse completamente en su interior (Figura 2.4).

4.2. Flujo de trabajo basado en secuenciación de tercera generación

4.2.1. Alineamiento

Las lecturas largas correspondientes a COLO829 (número de acceso del *European Nucleotide Archive*, ERX2765498) se alinearon frente al genoma humano de referencia en su versión GRCh37 con el alineador *Minimap2* (versión 2.17), dando como resultado un archivo con los alineamientos en formato SAM, los cuales fueron ordenados y almacenados en un archivo comprimido (formato BAM) por el programa *SAMtools* (comando *sort*).

4.2.2. Identificación de variantes

Se ejecutó *Sniffles* (versión 1.0.12) (Sedlazeck et al., 2018) sobre el archivo de alineamientos de COLO829. La ausencia de datos de secuenciación de tercera generación de un tejido sano

del mismo paciente impide determinar si las variantes identificadas son somáticas o germinales. Además, *Sniffles* no presenta un límite de longitud máximo para las variantes identificadas, pero sí uno de longitud mínima que, por defecto, es de 30 nucleótidos. Esto se debe a que uno de los inconvenientes de las plataformas de ONT, como *MinION*, es la existencia de un 2 - 13 % de error en la determinación de las bases que componen las lecturas (Logsdon et al., 2020). De tal forma, cuanto menor sea el límite inferior, existe una mayor probabilidad de identificar variantes no reales (falsos positivos). En palabras de uno de los desarrolladores de *Sniffles*, el programa se ha conseguido ejecutar correctamente hasta un límite de longitud mínimo de 10 nucleótidos, por debajo del cual no recomiendan trabajar. En nuestro caso, se procedió a ejecutar de forma preliminar *Sniffles* sobre el archivo de alineamientos con las opciones por defecto.

4.2.3. Análisis de las variantes identificadas

Materiales y métodos 4.1.5.

4.3. Recursos

Los flujos de trabajo creados en este proyecto se encuentran alojados en un repositorio de *GitLab* (`https://gitlab.com/amartin97/cancer_sv`). Los archivos con los tres análisis realizados (flujo de trabajo basado en lecturas cortas, el basado en lecturas largas y su comparación) se encuentran alojados en otro repositorio (`https://gitlab.com/amartin97/cancer_sv_secondary`). La ejecución de cada uno de los programas integrantes de los flujos de trabajo se llevó a cabo haciendo uso del clúster mantenido por la Unidad de Bioinformática del CNIO (`https://bu_cnio.gitlab.io/hpc/cluster/`; un nodo de computación básico de 24 CPUs y 32 Gigabytes de memoria RAM, y otro con 224 CPUs y 2 Terabytes de memoria RAM), con acceso remoto vía ssh. Sin embargo, la escritura de los flujos de trabajo, así como del código empleado en los análisis posteriores se ha realizado en un equipo con procesador Intel(R) Core(TM) i5-8250U (4 Núcleos, 6M Cache, 1.6GHz hasta 3.4GHz) con 8 Gigabytes de memoria RAM bajo el sistema operativo Ubuntu 20.04.2 LTS.

Glosario de acrónimos

- **ADN**: Ácido desoxirribonucleico
- **ARN**: Ácido ribonucleico
- **ARNm**: Ácido ribonucleico mensajero
- **BAM**: *Binary Alignment Map*
- **BWA**: *Burrows-Wheeler Aligner*
- **CNV**: *Copy Number Variant*
- **DEL**: deleción
- **DUP**: duplicación
- **EVS**: *Empirical Variant Score*
- **GSC**: *Canada's Michael Smith Genome Sciences Centre*
- **HPC**: *High Performance Computing*
- **indel**: inserción-deleción
- **INS**: inserción
- **INV**: inversión
- **ONT**: *Oxford Nanopore Technologies*
- **PCR**: *Polymerase Chain Reaction*
- **SAM**: *Sequence Alignment Map*
- **SNV**: *Single Nucleotide Variant*
- **SV**: *Structural Variant*
- **TGen**: *Translational Genomics Research Institute*
- **UCSC**: Universidad de California en Santa Cruz
- **VCF**: *Variant Calling Format*

Bibliografía

Haley J. Abel, David E. Larson, Allison A. Regier, Colby Chiang, Indraniel Das, Krishna L. Kanchi, Ryan M. Layer, Benjamin M. Neale, William J. Salerno, Catherine Reeves, Steven Buyske, Tara C. Matise, Donna M. Muzny, Michael C. Zody, Eric S. Lander, Susan K. Dutcher, Nathan O. Stitziel, and Ira M. Hall. Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, 583(7814):83–89, July 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2371-0. URL https://www.nature.com/articles/s41586-020-2371-0.

Rehan Akbani, Kadir C. Akdemir, B. Arman Aksoy, Monique Albert, Adrian Ally, Samirkumar B. Amin, Harindra Arachchi, Arshi Arora, J. Todd Auman, Brenda Ayala, Julien Baboud, Miruna Balasundaram, Saianand Balu, Nandita Barnabas, John Bartlett, Pam Bartlett, Boris C. Bastian, Stephen B. Baylin, Madhusmita Behera, Dmitry Belyaev, Christopher Benz, Brady Bernard, Rameen Beroukhim, Natalie Bir, Aaron D. Black, Tom Bodenheimer, Lori Boice, Genevieve M. Boland, Riccardo Bono, Moiz S. Bootwalla, Marcus Bosenberg, Jay Bowen, Reanne Bowlby, Christopher A. Bristow, Laura Brockway-Lunardi, Denise Brooks, Jakub Brzezinski, Wiam Bshara, Elizabeth Buda, William R. Burns, Yaron S. N. Butterfield, Michael Button, Tiffany Calderone, Giancarlo Antonini Cappellini, Candace Carter, Scott L. Carter, Lynn Cherney, Andrew D. Cherniack, Aaron Chevalier, Lynda Chin, Juok Cho, Raymond J. Cho, Yoon-La Choi, Andy Chu, Sudha Chudamani, Kristian Cibulskis, Giovanni Ciriello, Amanda Clarke, Stephen Coons, Leslie Cope, Daniel Crain, Erin Curley, Ludmila Danilova, Stefania D'Atri, Tanja Davidsen, Michael A. Davies, Keith A. Delman, John A. Demchok, Qixia A. Deng, Yonathan Lissanu Deribe, Noreen Dhalla, Rajiv Dhir, Daniel DiCara, Michael Dinikin, Michael Dubina, J. Stephen Ebrom, Sophie Egea, Greg Eley, Jay Engel, Jennifer M. Eschbacher, Konstantin V. Fedosenko, Ina Felau, Timothy Fennell, Martin L. Ferguson, Sheila Fisher, Keith T. Flaherty, Scott Frazer, Jessica Frick, Victoria Fulidou, Stacey B. Gabriel, Jianjiong Gao, Johanna Gardner, Levi A. Garraway, Julie M. Gastier-Foster, Carmelo Gaudioso, Nils Gehlenborg, Giannicola Genovese, Mark Gerken, Jeffrey E. Gershenwald, Gad Getz, Carmen Gomez-Fernandez, Thomas Gribbin, Jonna Grimsby, Benjamin Gross, Ranabir Guin, Tony Gutschner, Angela Hadjipanayis, Ruth Halaban, Benjamin Hanf, David Haussler, Lauren E. Haydu, D. Neil Hayes, Nicholas K. Hayward, David I. Heiman, Lynn Herbert, James G. Herman, Peter Hersey, Katherine A. Hoadley, Eran Hodis, Robert A. Holt, Dave SB Hoon, Susan Hoppough, Alan P. Hoyle, Franklin W. Huang, Mei Huang, Sharon Huang, Carolyn M. Hutter, Matthew Ibbs, Lisa Iype, Anders Jacobsen, Valerie Jakrot, Alyssa Janning, William R. Jeck, Stuart R. Jefferys, Mark A. Jensen, Corbin D. Jones, Steven J. M. Jones, Zhenlin Ju, Hojabr Kakavand, Hyojin Kang, Richard F. Kefford, Fadlo R. Khuri, Jaegil Kim, John M. Kirkwood, Joachim Klode, Anil Korkut, Konstanty Korski, Michael Krauthammer, Raju Kucherlapati, Lawrence N. Kwong, Witold Kycler, Marc Ladanyi, Phillip H. Lai, Peter W. Laird, Eric Lander, Michael S. Lawrence, Alexander J. Lazar, Radoslaw Łaźniak, Darlene Lee, Jeffrey E. Lee, Junehawk Lee, Kenneth Lee, Semin Lee, William Lee, Ewa Leporowska, Kristen M. Leraas, Haiyan I. Li, Tara M. Lichtenberg, Lee Lichtenstein, Pei Lin, Shiyun Ling, Jia Liu, Ouida Liu, Wenbin Liu, Georgina V. Long, Yiling Lu, Singer Ma, Yussanne Ma, Andrzej Mackiewicz, Harshad S. Mahadeshwar, Jared Malke, David Mallery, Georgy M. Manikhas, Graham J. Mann, Marco A. Marra, Brenna Matejka, Michael Mayo, Sousan Mehrabi, Shaowu Meng, Matthew Meyerson, Piotr A. Miecz-

kowski, John P. Miller, Martin L. Miller, Gordon B. Mills, Fedor Moiseenko, Richard A. Moore, Scott Morris, Carl Morrison, Donald Morton, Stergios Moschos, Lisle E. Mose, Florian L. Muller, Andrew J. Mungall, Dawid Murawa, Pawel Murawa, Bradley A. Murray, Luigi Nezi, Sam Ng, Dana Nicholson, Michael S. Noble, Adeboye Osunkoya, Taofeek K. Owonikoko, Bradley A. Ozenberger, Elena Pagani, Oxana V. Paklina, Angeliki Pantazi, Michael Parfenov, Jeremy Parfitt, Peter J. Park, Woong-Yang Park, Joel S. Parker, Francesca Passarelli, Robert Penny, Charles M. Perou, Todd D. Pihl, Olga Potapova, Victor G. Prieto, Alexei Protopopov, Michael J. Quinn, Amie Radenbaugh, Kunal Rai, Suresh S. Ramalingam, Ayush T. Raman, Nilsa C. Ramirez, Ricardo Ramirez, Uma Rao, W. Kimryn Rathmell, Xiaojia Ren, Sheila M. Reynolds, Jeffrey Roach, A. Gordon Robertson, Merrick I. Ross, Jason Roszik, Giandomenico Russo, Gordon Saksena, Charles Saller, Yardena Samuels, Chris Sander, Cindy Sander, George Sandusky, Netty Santoso, Melissa Saul, Robyn PM Saw, Dirk Schadendorf, Jacqueline E. Schein, Nikolaus Schultz, Steven E. Schumacher, Charles Schwallier, Richard A. Scolyer, Jonathan Seidman, Pedamallu Chandra Sekhar, Harmanjatinder S. Sekhon, Yasin Senbabaoglu, Sahil Seth, Kerwin F. Shannon, Samantha Sharpe, Norman E. Sharpless, Kenna R. Mills Shaw, Candace Shelton, Troy Shelton, Ronglai Shen, Margi Sheth, Yan Shi, Carolyn J. Shiau, Ilya Shmulevich, Gabriel L. Sica, Janae V. Simons, Rileen Sinha, Payal Sipahimalani, Heidi J. Sofia, Matthew G. Soloway, Xingzhi Song, Carrie Sougnez, Andrew J. Spillane, Arkadiusz Spychała, Jonathan R. Stretch, Joshua Stuart, Wiktoria M. Suchorska, Antje Sucker, S. Onur Sumer, Yichao Sun, Maria Synott, Barbara Tabak, Teresa R. Tabler, Angela Tam, Donghui Tan, Jiabin Tang, Roy Tarnuzzer, Katherine Tarvin, Honorata Tatka, Barry S. Taylor, Marek Teresiak, Nina Thiessen, John F. Thompson, Leigh Thorne, Vesteinn Thorsson, Jeffrey M. Trent, Timothy J. Triche, Kenneth Y. Tsai, Peiling Tsou, David J. Van Den Berg, Eliezer M. Van Allen, Umadevi Veluvolu, Roeland G. Verhaak, Douglas Voet, Olga Voronina, Vonn Walter, Jessica S. Walton, Yunhu Wan, Yuling Wang, Zhining Wang, Scot Waring, Ian R. Watson, Nils Weinhold, John N. Weinstein, Daniel J. Weisenberger, Peter White, Matthew D. Wilkerson, James S. Wilmott, Lisa Wise, Maciej Wiznerowicz, Scott E. Woodman, Chang-Jiun Wu, Chia-Chin Wu, Junyuan Wu, Ye Wu, Ruibin Xi, Andrew Wei Xu, Da Yang, Liming Yang, Lixing Yang, Travis I. Zack, Jean C. Zenklusen, Hailei Zhang, Jianhua Zhang, Wei Zhang, Xiaobei Zhao, Jingchun Zhu, Kelsey Zhu, Lisa Zimmer, Erik Zmuda, and Lihua Zou. Genomic Classification of Cutaneous Melanoma. *Cell*, 161(7): 1681–1696, June 2015. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2015.05.044. URL https://www.cell.com/cell/abstract/S0092-8674(15)00634-0.

Adam Ameur, Huiwen Che, Marcel Martin, Ignas Bunikis, Johan Dahlberg, Ida Höijer, Susana Häggqvist, Francesco Vezzi, Jessica Nordlund, Pall Olason, Lars Feuk, and Ulf Gyllensten. De Novo Assembly of Two Swedish Genomes Reveals Missing Segments from the Human GRCh38 Reference and Improves Variant Calling of Population-Scale Sequencing Data. *Genes*, 9(10): 486, October 2018. doi: 10.3390/genes9100486. URL https://www.mdpi.com/2073-4425/9/10/486.

Simon Andrews. FastQC: A quality control tool for high throughput sequence data, 2019. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Mark J. P. Chaisson, Richard K. Wilson, and Evan E. Eichler. Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*, 16(11):627–640, November 2015. ISSN 1471-0064. doi: 10.1038/nrg3933. URL https://www.nature.com/articles/nrg3933.

Mark J. P. Chaisson, Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, Oscar L. Rodriguez, Li Guo, Ryan L. Collins, Xian Fan, Jia Wen, Robert E. Handsaker, Susan Fairley, Zev N. Kronenberg, Xiangmeng Kong, Fereydoun Hormozdiari, Dillon Lee, Aaron M. Wenger, Alex R. Hastie, Danny Antaki, Thomas Anantharaman, Peter A. Audano, Harrison Brand, Stuart Cantsilieris, Han Cao, Eliza Cerveira, Chong

Chen, Xintong Chen, Chen-Shan Chin, Zechen Chong, Nelson T. Chuang, Christine C. Lambert, Deanna M. Church, Laura Clarke, Andrew Farrell, Joey Flores, Timur Galeev, David U. Gorkin, Madhusudan Gujral, Victor Guryev, William Haynes Heaton, Jonas Korlach, Sushant Kumar, Jee Young Kwon, Ernest T. Lam, Jong Eun Lee, Joyce Lee, Wan-Ping Lee, Sau Peng Lee, Shantao Li, Patrick Marks, Karine Viaud-Martinez, Sascha Meiers, Katherine M. Munson, Fabio C. P. Navarro, Bradley J. Nelson, Conor Nodzak, Amina Noor, Sofia Kyriazopoulou-Panagiotopoulou, Andy W. C. Pang, Yunjiang Qiu, Gabriel Rosanio, Mallory Ryan, Adrian Stütz, Diana C. J. Spierings, Alistair Ward, AnneMarie E. Welch, Ming Xiao, Wei Xu, Chengsheng Zhang, Qihui Zhu, Xiangqun Zheng-Bradley, Ernesto Lowy, Sergei Yakneen, Steven McCarroll, Goo Jun, Li Ding, Chong Lek Koh, Bing Ren, Paul Flicek, Ken Chen, Mark B. Gerstein, Pui-Yan Kwok, Peter M. Lansdorp, Gabor T. Marth, Jonathan Sebat, Xinghua Shi, Ali Bashir, Kai Ye, Scott E. Devine, Michael E. Talkowski, Ryan E. Mills, Tobias Marschall, Jan O. Korbel, Evan E. Eichler, and Charles Lee. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10(1):1784, April 2019. ISSN 2041-1723. doi: 10.1038/s41467-018-08148-z. URL https://www.nature.com/articles/s41467-018-08148-z.

Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222, April 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv710. URL https://doi.org/10.1093/bioinformatics/btv710.

George W. Cook, Michael G. Benton, Wallace Akerley, George F. Mayhew, Cynthia Moehlenkamp, Denise Raterman, Daniel L. Burgess, William J. Rowell, Christine Lambert, Kevin Eng, Jenny Gu, Primo Baybayan, John T. Fussell, Heath D. Herbold, John M. O'Shea, Thomas K. Varghese, and Lyska L. Emerson. Structural variation and its potential impact on genome instability: Novel discoveries in the EGFR landscape by long-read sequencing. *PLOS ONE*, 15(1):e0226340, January 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0226340. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0226340.

David W. Craig, Sara Nasser, Richard Corbett, Simon K. Chan, Lisa Murray, Christophe Legendre, Waibhav Tembe, Jonathan Adkins, Nancy Kim, Shukmei Wong, Angela Baker, Daniel Enriquez, Stephanie Pond, Erin Pleasance, Andrew J. Mungall, Richard A. Moore, Timothy McDaniel, Yussanne Ma, Steven J. M. Jones, Marco A. Marra, John D. Carpten, and Winnie S. Liang. A somatic reference standard for cancer genome sequencing. *Scientific Reports*, 6(1):1–11, April 2016. ISSN 2045-2322. doi: 10.1038/srep24607. URL https://www.nature.com/articles/srep24607.

Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27 (15):2156–2158, August 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr330. URL https://doi.org/10.1093/bioinformatics/btr330.

Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, and Christine Van Broeckhoven. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15):2666–2669, August 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty149. URL https://doi.org/10.1093/bioinformatics/bty149.

Wouter De Coster, Peter De Rijk, Arne De Roeck, Tim De Pooter, Svenn D'Hert, Mojca Strazisar, Kristel Sleegers, and Christine Van Broeckhoven. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research*, 29

(7):1178–1187, July 2019. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.244939.118. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.244939.118.

Jawad Fares, Mohamad Y. Fares, Hussein H. Khachfe, Hamza A. Salhab, and Youssef Fares. Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduction and Targeted Therapy*, 5(1):1–17, March 2020. ISSN 2059-3635. doi: 10.1038/s41392-020-0134-x. URL https://www.nature.com/articles/s41392-020-0134-x.

Nazeefa Fatima, Anna Petri, Ulf Gyllensten, Lars Feuk, and Adam Ameur. Evaluation of Single-Molecule Sequencing Technologies for Structural Variant Detection in Two Swedish Human Genomes. *Genes*, 11(12):1444, November 2020. ISSN 2073-4425. doi: 10.3390/genes11121444. URL https://www.mdpi.com/2073-4425/11/12/1444.

Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, January 2019. ISSN 0305-1048. doi: 10.1093/nar/gky955. URL https://doi.org/10.1093/nar/gky955.

Free Software Foundation, Inc. GNU bash, 2019. URL http://www.gnu.org/software/bash/.

Sabrina Giglio, Karl W. Broman, Naomichi Matsumoto, Vladimiro Calvari, Giorgio Gimelli, Thomas Neumann, Hirofumi Ohashi, Lucille Voullaire, Daniela Larizza, Roberto Giorda, Jim L. Weber, David H. Ledbetter, and Orsetta Zuffardi. Olfactory Receptor–Gene Clusters, Genomic-Inversion Polymorphisms, and Common Chromosome Rearrangements. *The American Journal of Human Genetics*, 68(4):874–883, April 2001. ISSN 0002-9297, 1537-6605. doi: 10.1086/319506. URL https://www.cell.com/ajhg/abstract/S0002-9297(07)61414-1.

Tingting Gong, Vanessa M Hayes, and Eva K F Chan. Detection of somatic structural variants from short-read next-generation sequencing data. *Briefings in Bioinformatics*, page bbaa056, May 2020. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbaa056. URL https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa056/5831479.

Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, June 2016. ISSN 1471-0064. doi: 10.1038/nrg.2016.49. URL https://www.nature.com/articles/nrg.2016.49.

Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, and Johannes Köster. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7):475–476, July 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0046-7. URL https://www.nature.com/articles/s41592-018-0046-7.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan

Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2649-2. URL https://www.nature.com/articles/s41586-020-2649-2.

John Huddleston, Mark J. P. Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David Gordon, Tina A. Graves-Lindsay, Katherine M. Munson, Zev N. Kronenberg, Laura Vives, Paul Peluso, Matthew Boitano, Chen-Shin Chin, Jonas Korlach, Richard K. Wilson, and Evan E. Eichler. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*, 27(5):677–685, January 2017. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.214007.116. URL http://genome.cshlp.org/content/27/5/677.

John D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.55. URL http://ieeexplore.ieee.org/document/4160265/.

Illumina, Inc. Library Preparation Kits | Optimized for Illumina sequencers, 2021. URL https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits.html.

Jeff Reback, Wes McKinney, jbrockmendel, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, gfyoung, Simon Hawkins, Sinhrks, Matthew Roeschke, Adam Klein, Terji Petersen, Jeff Tratner, Chang She, William Ayd, Shahar Naveh, Marc Garcia, Jeremy Schendel, Andy Hayden, Daniel Saxton, patrick, Vytautas Jancauskas, Ali McMaster, Pietro Battiston, Skipper Seabold, Marco Gorelli, Kaiqi Dong, chris-b1, h-vetinari, and Stephan Hoyer. pandas-dev/pandas: Pandas 1.2.1, January 2021. URL https://zenodo.org/record/4452601.

Sangtae Kim, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, Yeonbin Kim, Doruk Beyter, Peter Krusche, and Christopher T. Saunders. Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, 15(8):591–594, August 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0051-x. URL https://www.nature.com/articles/s41592-018-0051-x.

Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter development team. Jupyter Notebooks – a publishing format for reproducible computational workflows. In Fernando Loizides and Birgit Scmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87–90. IOS Press, 2016. doi: 10.3233/978-1-61499-649-1-87. URL https://eprints.soton.ac.uk/403913/.

Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, October 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts480. URL https://doi.org/10.1093/bioinformatics/bts480.

Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. May 2013. URL http://arxiv.org/abs/1303.3997. arXiv: 1303.3997.

Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty191. URL https://doi.org/10.1093/bioinformatics/bty191.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August

2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352. URL https://doi.org/10.1093/bioinformatics/btp352.

E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293, October 2009. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1181369. URL https://www.sciencemag.org/lookup/doi/10.1126/science.1181369.

Glennis A. Logsdon, Mitchell R. Vollger, and Evan E. Eichler. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10):597–614, October 2020. ISSN 1471-0064. doi: 10.1038/s41576-020-0236-x. URL https://www.nature.com/articles/s41576-020-0236-x.

Alberto Magi, Davide Bolognini, Niccoló Bartalucci, Alessandra Mingrino, Roberto Semeraro, Luna Giovannini, Stefania Bonifacio, Daniela Parrini, Elisabetta Pelo, Francesco Mannelli, Paola Guglielmelli, and Alessandro Maria Vannucchi. Nano-GLADIATOR: real-time detection of copy number alterations from nanopore sequencing data. *Bioinformatics*, 35(21):4213–4221, November 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz241. URL https://doi.org/10.1093/bioinformatics/btz241.

H. G. Morse and G. E. Moore. Cytogenetic homogeneity in eight independent sites in a case of malignant melanoma. *Cancer Genetics and Cytogenetics*, 69(2):108–112, September 1993. ISSN 0165-4608. doi: 10.1016/0165-4608(93)90083-x.

Maria Nattestad, Sara Goodwin, Karen Ng, Timour Baslan, Fritz J. Sedlazeck, Philipp Rescheneder, Tyler Garvin, Han Fang, James Gurtowski, Elizabeth Hutton, Elizabeth Tseng, Chen-Shan Chin, Timothy Beck, Yogi Sundaravadanam, Melissa Kramer, Eric Antoniou, John D. McPherson, James Hicks, W. Richard McCombie, and Michael C. Schatz. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Research*, 28(8):1126–1135, January 2018. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.231100.117. URL http://genome.cshlp.org/content/28/8/1126.

Ching Ching Ng, Kumiko Koyama, Shu Okamura, Hisato Kondoh, Yoshiki Takei, and Yusuke Nakamura. Isolation and characterization of a novel TP53-inducible gene, TP53TG3. *Genes, Chromosomes and Cancer*, 26(4):329–335, 1999. ISSN 1098-2264. doi: https://doi.org/10.1002/(SICI)1098-2264(199912)26:4<329::AID-GCC7>3.0.CO;2-C. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291098-2264%28199912%2926%3A4%3C329%3A%3AAID-GCC7%3E3.0.CO%3B2-C.

Konstantin Okonechnikov, Ana Conesa, and Fernando García-Alcalde. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32(2):292–294, January 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv566. URL https://doi.org/10.1093/bioinformatics/btv566.

Oxford Nanopore Technologies. How it works, 2020. URL http://nanoporetech.com/how-it-works.

Brock A. Peters, Bahram G. Kermani, Andrew B. Sparks, Oleg Alferov, Peter Hong, Andrei Alexeev, Yuan Jiang, Fredrik Dahl, Y. Tom Tang, Juergen Haas, Kimberly Robasky, Alexander Wait Zaranek, Je-Hyuk Lee, Madeleine Price Ball, Joseph E. Peterson, Helena Perazich, George Yeung, Jia Liu, Linsu Chen, Michael I. Kennemer, Kaliprasad Pothuraju, Karel Konvicka, Mike Tsoupko-Sitnikov, Krishna P. Pant, Jessica C. Ebert, Geof-

frey B. Nilsen, Jonathan Baccash, Aaron L. Halpern, George M. Church, and Radoje Drmanac. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, 487(7406):190–195, July 2012. ISSN 1476-4687. doi: 10.1038/nature11236. URL https://www.nature.com/articles/nature11236.

Erin D. Pleasance, R. Keira Cheetham, Philip J. Stephens, David J. McBride, Sean J. Humphray, Chris D. Greenman, Ignacio Varela, Meng-Lay Lin, Gonzalo R. Ordóñez, Graham R. Bignell, Kai Ye, Julie Alipaz, Markus J. Bauer, David Beare, Adam Butler, Richard J. Carter, Lina Chen, Anthony J. Cox, Sarah Edkins, Paula I. Kokko-Gonzales, Niall A. Gormley, Russell J. Grocock, Christian D. Haudenschild, Matthew M. Hims, Terena James, Mingming Jia, Zoya Kingsbury, Catherine Leroy, John Marshall, Andrew Menzies, Laura J. Mudie, Zemin Ning, Tom Royce, Ole B. Schulz-Trieglaff, Anastassia Spiridou, Lucy A. Stebbings, Lukasz Szajkowski, Jon Teague, David Williamson, Lynda Chin, Mark T. Ross, Peter J. Campbell, David R. Bentley, P. Andrew Futreal, and Michael R. Stratton. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463 (7278):191–196, January 2010. ISSN 1476-4687. doi: 10.1038/nature08658. URL https://www.nature.com/articles/nature08658.

Balananda-Dhurjati Kumar Putcha, Xu Jia, Venkat Rao Katkoori, Chura Salih, Chandrakumar Shanmugam, Trafina Jadhav, Liselle C. Bovell, Michael P. Behring, Tom Callens, Ludwine Messiaen, Sejong Bae, William E. Grizzle, Karan P. Singh, and Upender Manne. Clinical Implications of Rabphillin-3A-Like Gene Alterations in Breast Cancer. *PLOS ONE*, 10(6): e0129216, June 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0129216. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0129216.

R Core Team. R: A Language and Environment for Statistical Computing, 2020. URL https://www.R-project.org.

Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, September 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts378. URL https://doi.org/10.1093/bioinformatics/bts378.

E. Premkumar Reddy, Roberta K. Reynolds, Eugenio Santos, and Mariano Barbacid. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature*, 300(5888):149–152, November 1982. ISSN 1476-4687. doi: 10.1038/300149a0. URL https://www.nature.com/articles/300149a0.

Fritz J. Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C. Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6):461–468, June 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0001-7. URL https://www.nature.com/articles/s41592-018-0001-7.

A.F.A. Smith, R. Hubley, and P. Green. RepeatMasker Open-4.0, 2015. URL http://www.repeatmasker.org/.

Zbyslaw Sondka, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, November 2018. ISSN 1474-1768. doi: 10.1038/s41568-018-0060-1. URL https://www.nature.com/articles/s41568-018-0060-1.

Pieter Spealman, Jaden Burrell, and David Gresham. Inverted duplicate DNA sequences increase translocation rates through sequencing nanopores resulting in reduced base calling accuracy. *Nucleic Acids Research*, 48(9):4940–4945, May 2020. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkaa206. URL https://academic.oup.com/nar/article/48/9/4940/5816855.

Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, April 2009. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature07943. URL http://www.nature.com/articles/nature07943.

Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, Miriam K. Konkel, Ankit Malhotra, Adrian M. Stütz, Xinghua Shi, Francesco Paolo Casale, Jieming Chen, Fereydoun Hormozdiari, Gargi Dayama, Ken Chen, Maika Malig, Mark J. P. Chaisson, Klaudia Walter, Sascha Meiers, Seva Kashin, Erik Garrison, Adam Auton, Hugo Y. K. Lam, Xinmeng Jasmine Mu, Can Alkan, Danny Antaki, Taejeong Bae, Eliza Cerveira, Peter Chines, Zechen Chong, Laura Clarke, Elif Dal, Li Ding, Sarah Emery, Xian Fan, Madhusudan Gujral, Fatma Kahveci, Jeffrey M. Kidd, Yu Kong, Eric-Wubbo Lameijer, Shane McCarthy, Paul Flicek, Richard A. Gibbs, Gabor Marth, Christopher E. Mason, Androniki Menelaou, Donna M. Muzny, Bradley J. Nelson, Amina Noor, Nicholas F. Parrish, Matthew Pendleton, Andrew Quitadamo, Benjamin Raeder, Eric E. Schadt, Mallory Romanovitch, Andreas Schlattl, Robert Sebra, Andrey A. Shabalin, Andreas Untergasser, Jerilyn A. Walker, Min Wang, Fuli Yu, Chengsheng Zhang, Jing Zhang, Xiangqun Zheng-Bradley, Wanding Zhou, Thomas Zichner, Jonathan Sebat, Mark A. Batzer, Steven A. McCarroll, Ryan E. Mills, Mark B. Gerstein, Ali Bashir, Oliver Stegle, Scott E. Devine, Charles Lee, Evan E. Eichler, and Jan O. Korbel. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, October 2015. ISSN 1476-4687. doi: 10.1038/nature15394. URL https://www.nature.com/articles/nature15394.

Srinath Sundararajan, Aye M. Thida, and Talel Badri. Metastatic Melanoma. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2020. URL http://www.ncbi.nlm.nih.gov/books/NBK470358/.

Clifford J. Tabin, Scott M. Bradley, Cornelia I. Bargmann, Robert A. Weinberg, Alex G. Papageorge, Edward M. Scolnick, Ravi Dhar, Douglas R. Lowy, and Esther H. Chang. Mechanism of activation of a human oncogene. *Nature*, 300(5888):143–149, November 1982. ISSN 1476-4687. doi: 10.1038/300143a0. URL https://www.nature.com/articles/300143a0.

Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31(12):2032–2034, June 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv098. URL https://doi.org/10.1093/bioinformatics/btv098.

Jose Espejo Valle-Inclan, Christina Stangl, Anouk C. de Jong, Lisanne F. van Dessel, Markus J. van Roosmalen, Jean C. A. Helmijr, Ivo Renkens, Sam de Blank, Chris J. de Witte, John W. M. Martens, Maurice P. H. M. Jansen, Martijn P. Lolkema, and Wigard P. Kloosterman. Rapid identification of genomic structural variations with nanopore sequencing enables blood-based cancer monitoring. *medRxiv*, page 19011932, November 2019. doi: 10.1101/19011932. URL https://www.medrxiv.org/content/10.1101/19011932v1.

Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

Xiaofei Yang, Wan-Ping Lee, Kai Ye, and Charles Lee. One reference genome is not enough. *Genome Biology*, 20(1):104, May 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1717-0. URL https://doi.org/10.1186/s13059-019-1717-0.

Grace X. Y. Zheng, Billy T. Lau, Michael Schnall-Levin, Mirna Jarosz, John M. Bell, Christopher M. Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A. Masquelier, Landon Merrill, Jessica M. Terry, Patrice A. Mudivarti, Paul W. Wyatt, Rajiv Bharadwaj, Anthony J.

Makarewicz, Yuan Li, Phillip Belgrader, Andrew D. Price, Adam J. Lowe, Patrick Marks, Gerard M. Vurens, Paul Hardenbol, Luz Montesclaros, Melissa Luo, Lawrence Greenfield, Alexander Wong, David E. Birch, Steven W. Short, Keith P. Bjornson, Pranav Patel, Erik S. Hopmans, Christina Wood, Sukhvinder Kaur, Glenn K. Lockwood, David Stafford, Joshua P. Delaney, Indira Wu, Heather S. Ordonez, Susan M. Grimes, Stephanie Greer, Josephine Y. Lee, Kamila Belhocine, Kristina M. Giorda, William H. Heaton, Geoffrey P. McDermott, Zachary W. Bent, Francesca Meschi, Nikola O. Kondov, Ryan Wilson, Jorge A. Bernate, Shawn Gauby, Alex Kindwall, Clara Bermejo, Adrian N. Fehr, Adrian Chan, Serge Saxonov, Kevin D. Ness, Benjamin J. Hindson, and Hanlee P. Ji. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34(3):303–311, March 2016. ISSN 1546-1696. doi: 10.1038/nbt.3432. URL https://www.nature.com/articles/nbt.3432.

A

Material suplementario

A.1. Figura suplementaria 1 (Figura A.1)

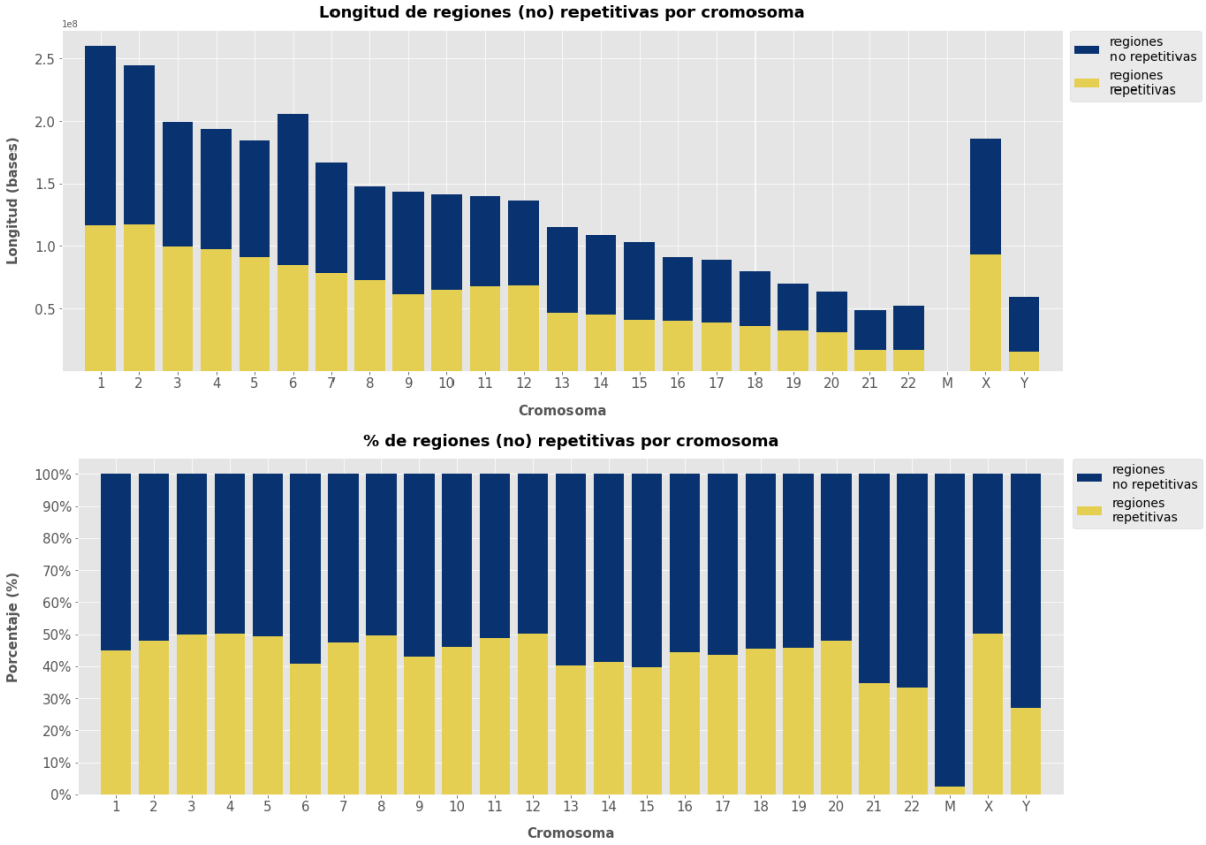

Figura A.1: Gráficos de barras apiladas representando, para cada cromosoma, el número (#, panel superior) o porcentaje (%, panel inferior) de nucleótidos en regiones no repetitivas en azul y el de aquellos en repetitivas en amarillo.

A.2. Figura suplementaria 2 (Figura A.2)

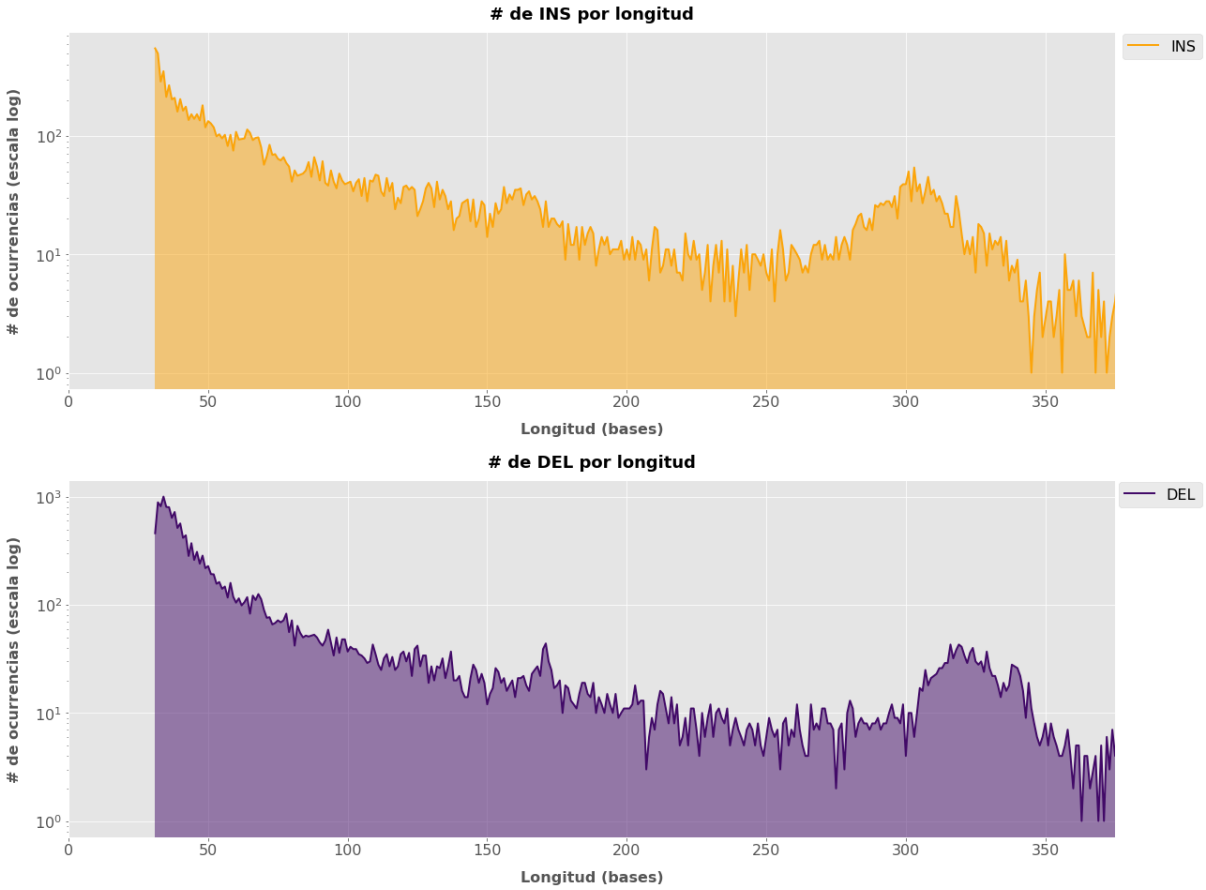

Figura A.2: Número (#), en escala logarítmica, de variantes por cada valor longitud posible, considerando el intervalo de longitudes desde la mínima a una de 375 nucleótidos. En ambos paneles, el conjunto empleado corresponde a todas las variantes resultantes del flujo de trabajo de lecturas largas. En el panel superior, las variantes representadas pertenecen al tipo inserción. En el panel inferior, al tipo deleción, mostrando los tamaños en valor absoluto.

A.3. Figura suplementaria 3 (Figura A.3)

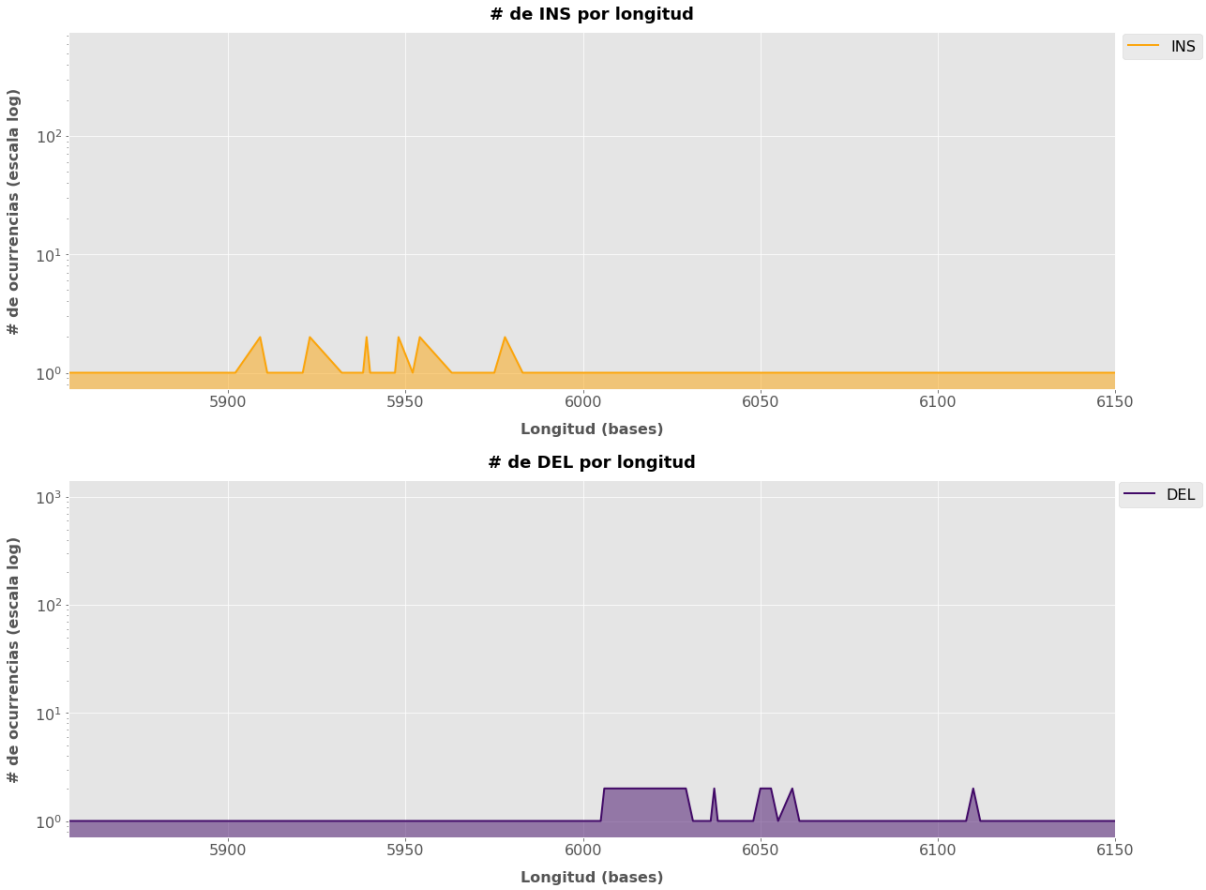

Figura A.3: Número (#), en escala logarítmica, de variantes por cada valor longitud posible, considerando el intervalo de longitudes desde 5 855 a 6 150 nucleótidos. En ambos paneles, el conjunto empleado corresponde a todas las variantes resultantes del flujo de trabajo de lecturas largas. En el panel superior, las variantes representadas pertenecen al tipo inserción. En el panel inferior, al tipo deleción, mostrando los tamaños en valor absoluto.

A.4. Tabla suplementaria 1 (Tabla A.1)

ID	Conjunto	*CHROM*	*POS*	*REF*	*ALT*
1	COLO829	chr11	18736573	GACT	G
	Strelka2		18736573	GACTGTGGGATACATACATACA	G
2	COLO829	chr12	38817159	TAGG	T
	Strelka2		38817159	TA	T

Tabla A.1: Variantes adicionalmente encontradas en el proceso de búsqueda de aquellas de la referencia somática de COLO829 (filas correspondientes al "Conjunto" COLO829) entre las identificadas por el flujo de trabajo de lecturas cortas (filas correspondientes al "Conjunto" Strelka2). Las columnas incluyen un identificador (ID) único para cada variante, el conjunto al que pertenece la expuesta en la fila correspondiente y la información de los campos *CHROM*, *POS*, *REF* y *ALT* del archivo de formato VCF al que pertenece la misma.

A.5. Tabla suplementaria 2 (Tabla A.2)

Resumen general	
# de lecturas	18 687 976
Longitud media de lectura	10 492.8
Calidad media de lectura	9.2
Longitud mediana de lectura	7 018
Calidad mediana de lectura	9.6
# de bases	196 088 528 773
#, porcentaje y megabases de lecturas por encima de los límites de calidad	
Q5	15 956 044 (85.4 %) 192856.6Mb
Q7	14 541 660 (77.8 %) 182650.7Mb
Q10	8 250 844 (44.2 %) 114028.4Mb
Q12	4 158 410 (22.3 %) 60774.7Mb
Q15	14 875 (0.1 %) 105.2Mb
Los 5 scores de calidad media más altos y sus longitudes de lectura	
21.3 (22)	
20.1 (32)	
20.0 (14)	
18.8 (15)	
18.7 (19)	
Las 5 lecturas más largas y su puntuación de calidad media	
1376732 (3.4)	
892310 (4.0)	
791615 (3.9)	
761898 (3.9)	
577781 (4.1)	

Tabla A.2: Resumen generado por el control de calidad ejecutado con *NanoPlot* sobre el archivo en formato FASTQ de lecturas largas de COLO829.

A.6. Tabla suplementaria 3 (Tabla A.3)

ID	Conjunto	*CHROM*	*POS*	*REF*	*ALT*
3	COLO829	chr7	110835883	T	TATTTC
	Sniffles		110835824	N	AATATGAAATTTATTAAAACCAACACAAAATTT ATGAGAATCTCTTTTTAAACCTGGAAATATTCA GAGAAACTTTTTTACGTGGTTTTAATGAATTTC ATATTTTCATG
4	COLO829	chr8	132332724	TA	T
	Sniffles		132332234	AAAAAATAATATATATATATATATATATATATATATATAT	N

Tabla A.3: Variantes encontradas en el proceso de búsqueda de aquellas de la referencia somática de COLO829 (filas correspondientes al "Conjunto" COLO829) entre las identificadas por el flujo de trabajo de lecturas largas (filas correspondientes al "Conjunto" Sniffles), incluyendo las filtraciones pertinentes (Resultados 2.2.1). Las columnas incluyen un identificador (ID) único para cada variante, el conjunto al que pertenece la expuesta en la fila correspondiente y la información de los campos *CHROM*, *POS*, *REF* y *ALT* del archivo de formato VCF al que pertenece la misma.

A.7. Tabla suplementaria 4 (Tabla A.4)

ID	Conjunto	*CHROM*	*POS*	*REF*	*ALT*
5	Strelka2	chr1	118516186	**TGTTAGAGAGAAAGCTTAAATTATGCCACATGGACA**	T
	Sniffles		118516186	GCTTC**TGTTAGAGAGAAAGCTTAAATTATGCCACATGGACA**	N
6	Strelka2	chr12	60398423	T	TA
	Sniffles		60398423	N	TATATATATATAATATATATATTATATATATTTA TATATATTATATATATAATATATATATTTATATA TATAATATATATATATATATATAATATATATA TTATATATATTAATATATAATATATATTATATAT ATATAATATATATATTATATATAATATATATA
7	Strelka2	chr10	69687148	**TATATATATATATATATATATATATATATATATATATA**TA	T
	Sniffles		69687148	**ATATATATATATATATATATATATATATATATATATATA**	N
8	Strelka2	chr15	72217526	ATG	A
	Manta		72217526	**ATGTGTATATATACACATATATACATGTGTATATATA** CACATATATACACA	ACG
	Sniffles		72217526	**ATGTGTATATATACACATATATACATGTGTATATATA**	N
9	COLO829	chr7	110835883	T	TATTTC
	Strelka2		110835883	T	TATTTC
	Manta		110835823	A	**AATATGAAATTTATTAAAACCAACACAAAATTT ATGAGAATCTCTTTTTAAACCTGGAAATATTCA GAGAAACTTTTTTACGTGGTTTTAATGAATTTC ATATTTTCATG**
	Sniffles		110835824	N	**AATATGAAATTTATTAAAACCAACACAAAATTT ATGAGAATCTCTTTTTAAACCTGGAAATATTCA GAGAAACTTTTTTACGTGGTTTTAATGAATTTC ATATTTTCATG**
10	COLO829	chr8	132332724	TA	T
	Strelka2		132332724	TA	T
	Manta		132332268	**ATAT**ATATATATATATATATATATATGTATAT	AG
	Sniffles		132332234	AAAAAATAATATATATATATATATATATATATATATATAT**ATAT**	N

Tabla A.4: Variantes encontradas en el proceso de búsqueda de aquellas de la referencia somática de COLO829 (filas correspondientes al "Conjunto" COLO829) entre las identificadas por el flujo de trabajo de lecturas cortas (filas correspondientes al "Conjunto" Strelka2), las identificadas por el de lecturas cortas hasta la ejecución de Manta (filas correspondientes al "Conjunto" Manta) y las identificadas por el de lecturas largas (filas correspondientes al "Conjunto" Sniffles), incluyendo las filtraciones pertinentes en cada caso (Resultados 2.3). Las columnas incluyen un identificador (ID) único para cada variante, el conjunto al que pertenece la expuesta en la fila correspondiente y la información de los campos *CHROM*, *POS*, *REF* y *ALT* del archivo de formato VCF al que pertenece la misma. Las secuencias resaltadas en amarillo o negrita corresponden a aquellas de algún modo coincidentes entre la variante en la referencia, la identificada por el flujo de trabajo de lecturas cortas (hasta la ejecución de *Manta* o de *Strelka2*) y la identificada por el de lecturas largas.